State-space modelling for high frequency data

Dordonnat, V.

2009

document version
Publisher's PDF, also known as Version of record

State-space modelling for high frequency data
Three applications to French national electricity load

VRIJE UNIVERSITEIT

State-space modelling for high frequency data
Three applications to French national electricity load

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Economische Wetenschappen en Bedrijfskunde
op vrijdag 4 september 2009 om 10.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Virginie Dordonnat

geboren te Dourdan, Frankrijk

promotor: prof.dr. S.J. Koopman
copromotor: dr. M. Ooms

Acknowledgements

Until the last month, I was not sure that I would succeed in finishing this dissertation but I finally did. Many people contributed to this achievement. I first want to thank the members of the jury who approved this doctoral thesis.

Writing a doctoral thesis was not part of the initial plan. I am thankful to the people at EDF R&D who suggested me to start a PhD project after having supervised me during my master's internship: Bertrand Vignal and Jean-Sebastien Roy. I also want to thank René Aid who also supported me in starting the project. I started the PhD project with Jerome Collet. I am happy to start working with him again. Thanks to Alain Dessertaine who accepted to continue the supervising of my work on the EDF side.

My PhD project has been financed by the load forecasting group from OSIRIS department, EDF R&D, Clamart, France. I learnt a lot from everybody over there and also enjoyed coffee, tea and good food. I successively shared a room with Yannig and Jairo. You were in the first line during the difficult moments, thank you so much for supporting during the too many doubting moments. Nicolas, thanks for your expert comments on my empirical results and for your impersonations. Also thanks for "enjoying" maybe one of the worst movies ever at maybe one of the most difficult moment. The three of you, joined by Amandine and Aurélie, thanks for all the afterwork drinks.

I had great support from my VU supervisors: Marius, Siem Jan, thank you so much for welcoming me in the econometrics department. Thanks for all the positive comments and criticisms, thanks for all the time spent reading (and correcting) my papers. You offered me a great research opportunity which also allowed me to enjoy the Amsterdam experience. Also thanks for the Dutch translation of the thesis summary.

I met a lot of nice people in Amsterdam, I am glad that I shared a room with them: Irma, Suncica, Brian, Borus, Taoying and Bahar. Thanks for all the dinners and discussions, and for being so friendly with the French girl coming once in a while to work at the VU.

I never would have succeeded without friends' support. Ahlame, you believed that I

would finish the thesis more often than I did, thanks for your endless support and the chatting moments after work in Paris, or the visits in Amsterdam. Karen, François and Nicolas, thanks for all our losers evenings, playing poker and eating pizzas or watching new TV series ... eating pizzas.

Thanks to the chiropodists group for the parties, gaming nights on Mario Kart or Naruto and even for the funny descriptions of the statistician's work! Gweltaz, thanks again for the last year's "holidays" in Quiberon, this was the occasion to experiment statistical modelling by the sea.

Last but not least, I want to dedicate this work to my family and especially to Vincent who stood this long period with me.

Paris, May 2009.

Contents

Chapter 1

Introduction

The motivation of this dissertation is the modelling and forecasting of high frequency data such as hourly electricity loads within the state-space framework. This chapter starts with a description of the short-term load forecasting issue. We illustrate the discussion with French national data. A review of alternative model specifications is presented next. Operational models in use at Electricité de France[1] are detailed as well as the motivation for new modelling with state-space methods. The second part of the chapter is an introduction to state-space modelling. Model specification and parameter estimation are discussed, we also describe practical implementation. This introduction chapter ends with an overview of the thesis and the description of the datasets used in our empirical studies.

1.1 Short-term electricity load forecasting

1.1.1 The need for short-term forecasting models

Accurate short-term forecasts for electricity loads are required for physical and market organization reasons. Producing accurate forecasts is most importantly required to ensure the permanent balance between electricity production and demand on the network. Except during black-out situations, the balance is effective, hence we can refer to electricity load or demand. Another reason is the market liberalization: all market participants must provide the national grid with demand forecasts for their own portfolio. In case of large forecast errors they have to face important financial penalties. Finally, market liberalization also brought up trading possibilities.

[1] EDF is the major electricity producer and retailer in France, www.edf.com

Short-term forecasts usually correspond to a forecasting horizon from the next (half-) hour up to the next week or even 10 days. Short-term forecasting models are therefore based on high frequency data, usually hourly or half-hourly measured. Martin (1999) presented a model for French demand measured every five minutes, arguing it was the optimal sampling interval to properly represent the daily load curve (corresponding to the load measured at each moment of the day). Taylor (2008) even suggested a forecasting model for data measured every minute to forecast from 10 minutes up to 30 minutes ahead.

Longer forecasting horizons are used for maintenance scheduling and investment planning (for network extension as well as producing capabilities). Mid-term forecasting (usually from the next month up to several years) models can have similar specifications to short-term forecasting models while long-term forecasting models (horizon > 5 years) imply economic modelling and studies for total annual energy. Long-term peak load is also forecast for specific peak production utilities investment planning. More generally, long-term forecasting for the load curve is required for an optimal production mix setting.

1.1.2 Description of French electricity demand

Hourly electricity demand is a typical example of high frequency data. We illustrate the main features of French national demand with the following figures. Figure 1.1(a) draws the daily mean of electricity demand, measured in MegaWatts (MW), on the (long) period January 1^{st}, 1997 until August 31^{st}, 2007. It shows that electricity demand is gradually increasing with economic growth so that an effective model requires a trend component. We also notice a repeating yearly pattern from one year to the following although this is not a stationary cycle as defined in time series theory. This distorted pattern also has to be taken into account for load modelling. Figure 1.1(b) focuses on year 2006 and we put electricity load in perspective with the daily mean of the national temperature in Celsius degrees (^oC) on the same period, presented in Figure 1.1(d). These pictures allow us to study the intrayearly pattern of the load in more detail. Both curves are following opposite directions in the winter: when the temperature decreases, demand for electric heating increases. During spring and fall, temperature has an almost null impact, the load mean level being stable during these periods. The temperature's intrayearly pattern does not explain all the load's intrayearly pattern: the remaining pattern could be explained by daylight intensity (although we do not have data for this) and economic activity. The impact of economic activity is clearly visible with the heavy fall of electricity demand in August, related to the summer holidays in France. Finally,

Figure 1.1(c) draws electricity daily mean demand against the daily mean of national temperature. The relationship is clearly non-linear. The three lines on the graph show how this non-linear curve can be approximated with three piecewise linear functions. The left line represents the heating effect with a negative slope, while the right one represents the cooling effect with a positive slope. The middle line is a horizontal one, illustrating the absence of temperature influence during mid-seasons. When extremely cold days days occur, an extra heating influence can be observed although this is not clearly visible here. The piecewise linear representation is adopted in the empirical applications of this thesis. A more sophisticated model can involve cubic splines for example. Note that temperature influence on demand is much larger during the winter and that only a few observations with warm temperatures (above 22°C) occur. Model specifications for the cooling effect should therefore be simpler than for the heating effect.

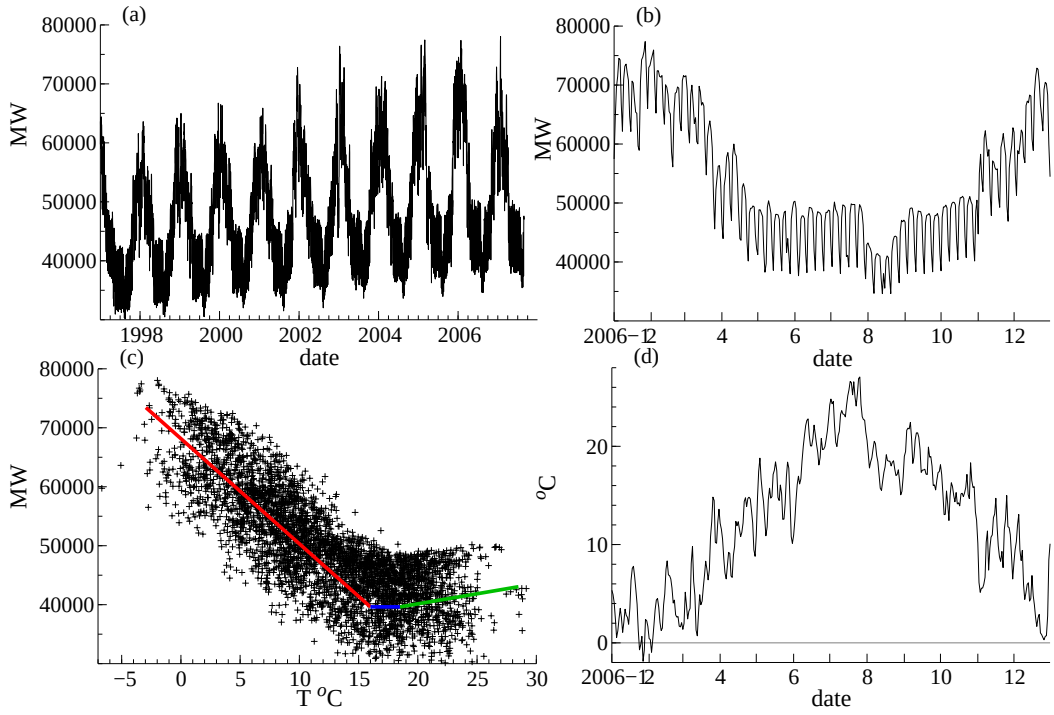

Figure 1.1: Daily mean of French national hourly electricity loads in MegaWatts: (a) from January 1^{st}, 1997 until August 31^{st}, 2007; (b) for year 2006; (c) against daily mean of French national temperature in Celsius degrees with piecewise linear adjustment (left, red: heating effect; center, blue: no temperature effect; right, green: cooling effect). Daily mean of French national temperature (d) for year 2006 .

Note that Figure 1.1 (except panel (d)) draws daily means of electricity demand so that the different features' description remains valid when working on electricity loads at a more aggregate level (daily or monthly loads).

Figure 1.2 draws hourly electricity loads for three months in 2006 showing the daily and weekly patterns of hourly electricity demand for different periods of the year: panel (a) corresponds to January (1^{st} day is a Sunday), panel (b) to May (1^{st} day is a Monday) and panel (c) to August (1^{st} day is a Tuesday). Weekdays and weekends can clearly be distinguished with a strong decrease of the load level during weekends as well as a different daily curve shape. This picture also allows to notice another particular feature of electricity demand due to bank holidays: e.g. May 1^{st} and 8^{th} correspond to Mondays, however the daily load curve is closer to the one of a regular weekend day than a regular working day; Thursday, 25^{th} of May is also a bank holiday so that the following Friday is a bridge day, and we can see clearly that the daily load curve is highly different from the other Fridays of the month. Finally, the three pictures also show the difference in the daily load curve between the different seasons: the evening peak is more pronounced in January.

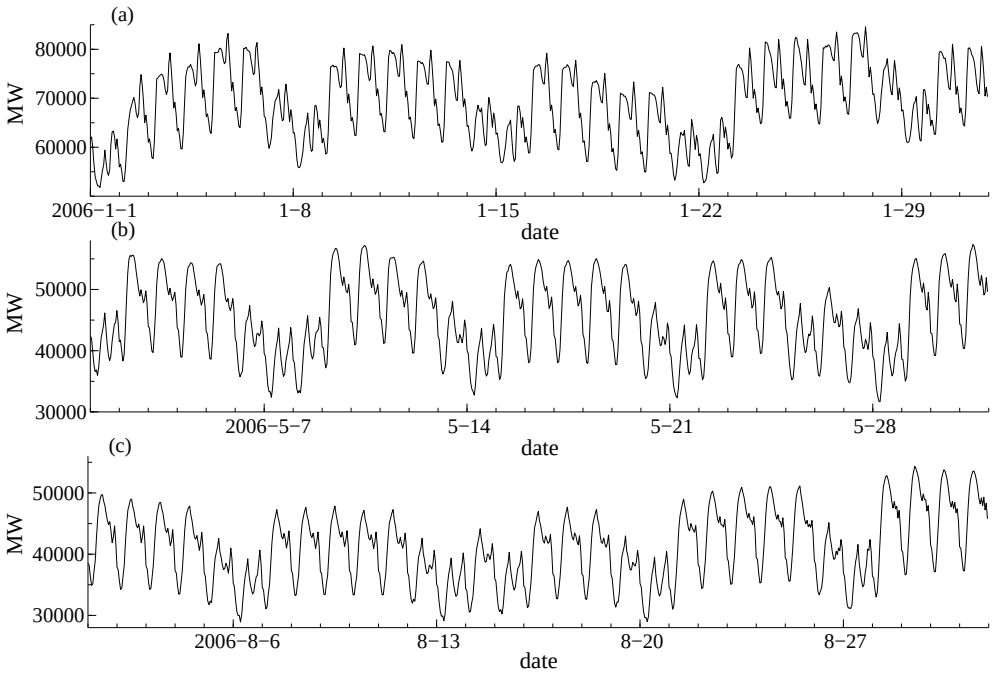

Figure 1.2: French national hourly electricity loads in MegaWatts: (a) in January 2006; (b) in May 2006; (c) in August 2006 .

Comparison with other countries We described the main features of French national electricity demand that have to be taken into account for the building of an effective modelling and forecasting strategy. These characteristics are in some way specific to this country and cannot be directly translated to other countries. It may therefore be complicated to suggest a general model for all electricity markets: particular adjust-

ments in the model specification are always required. For example, electric heating is of high importance in France so that the temperature effect in the winter is necessary for accurate forecasting, while electricity demand is less influenced in the summer due to the still relatively small use of air conditioning. The temperature-load curve for Spain described in Pardo, Meneua & Valor (2002) is much more symmetrical for low and high temperatures. The slope for the cooling is even more important than the one for heating on PJM-market data described in Sisworahardjo, El-Keib, Choi, Valenzuela, Brooks & El-Agtal (2006). The curve for South Australia (corrected for industrial demand) does not even have a negative slope for minimum temperatures but still remains non-linear, see Hyndman & Fan (2008). Peirson & Henley (1994) investigate thoroughly the relationship between electricity demand and temperature in Great Britain. They discuss the possibility to switch between lagged demand and temperature, and the construction of an effective approach to represent internal buildings temperature.

Bank holidays and vacation effects are also specific to the country under study. Moreover, due to the different calendar configurations from one year to another it is difficult to build a generic statistical methodology to properly forecast bank holidays, expert forecasts (or corrections to a model for regular days) are often more accurate.

1.1.3 Modelling strategies

Hourly (or half-hourly) electricity loads modelling and forecasting is a popular issue in academic research. Bunn & Farmer (1985) offer an early review while Lotufo & Minussi (1999) and Alfares & Nazeeruddin (2002) give recent updates. There are different possibilities to classify the different methodologies. We give as an indication, when available, the one-day ahead forecasting Mean Absolute Percentage Errors (MAPE) to illustrate the different accuracies reached by the models. Note that it is not our aim to compare the numbers themselves since forecasting accuracy strongly depends on the dataset.

Some authors build univariate time series models without exogenous variables. A thorough review of time series forecasting methods has been recently provided by De Gooijer & Hyndman (2006). Taylor (2003) adopted a double seasonal exponential smoothing for 8 training weeks of half-hourly British summer data in order to forecast 4 weeks from the next half-hour up to the next day (48 step-ahead forecasts) with a forecasting MAPE between 1.25 and 2%. Taylor, De Menezes & McSharry (2006) performed a comparative study between different univariate methods for Rio de Janeiro (MAPE >2.5%) and British data (MAPE between 1.25 and 2.25%) and Taylor & McSharry (2007) proposed

a benchmark on 10 European countries data, the overall MAPE for the different methods going from 1.5 up to 3%.

Other authors add exogenous variables: Harvey & Koopman (1993) used a structural time series model and included temperature using cubic splines to forecast Puget Sound Power and Light (PSPL) hourly loads, the forecasting MAPE was up to 3%. Their model was extended by Smith (2000) who built a Bayesian semi-parametric regression model on New South Wales data (forecasting MAPE $> 2\%$). Wang, Neng-ling, Hai-qing, Jian, Jia-dong & Liang-bo (2008) suggested a genetic algorithm for parameter estimation of an ARMAX model, applied to Chinese data. The model was estimated on only two months of summer data when the cooling effect is the most important, and forecasting accuracy was evaluated on only 9 days with a MAPE between 1 and 3.5%.

An alternative to univariate modelling is to build multiple-equation models, generally writing the same equation for each (half-)hour but with different parameter values. An hourly independent Two-Level Seasonal Autoregressive (TLSAR) model with no exogenous variables was suggested by Soares & Medeiros (2008) for Brazilian data and compared to benchmark models. Their MAPE was above 2.5% in most cases, including the special days in the study. Ramanathan, Engle, Granger, Vahid-Araghi & Brace (1997)) also built an independent model for each hour but included temperature effects, the overall forecasting MAPE for PSPL data was above 3%. Sisworahardjo et al. (2006) decomposed PJM electricity demand in a non-linear regression part related to temperature and a weather-independent component modelled as an ARIMA process. Data were split in different periods (weekdays/weekends, summer/winter) with a forecasting MAPE of 2.85%. Dependent models within the Bayesian framework were constructed by Smith & Kohn (2002) and Cottet & Smith (2003) for New South Wales electricity load, the former focused on a short dataset (3 weeks) and the latter obtained a forecasting MAPE for one year of data of 2.27%.

Artificial Neural Networks (ANN) for load forecasting were suggested by Hippert, Bunn & Souza (2005) with a post-sample MAPE around 2% while Support Vector Machines (SVM) for annual Taïwan data were proposed in Pai & Hong (2005).

Finally, some authors investigated semi or non-parametric modelling approaches: Poggi (1994) presented a nonparametric model for French half-hourly electricity loads where only regular days were considered and data corrected for weather effects. Kernel autoregression was used, reaching a 1.6% MAPE. De Gooijer & Ray (2003) built a multivariate regression model using POLYMARS on Australian data. More recently, a semi-parametric regression model for French national electricity load has been suggested by Lefieux (2007), reaching a forecasting MAPE of 0.9% for an April-July period, 1.8%

for a January-April period.

We should note that most of the different models described here do not consider special days: they are excluded from the study or corrected before model estimation to correspond to more general load behaviour. Furthermore, forecasting results are inhomogeneous. Accuracy is lower when working with local data (city, region) than with nation-wide data.

The inclusion of weather-based variables in electricity demand models is subject to discussion. If the temperature is changing in a smooth way, then it can be replaced by functions of the lagged load itself as argued by Taylor (2003). The explanatory variable used for short-term forecasting is therefore known instead of being subject to forecasting uncertainty, see also Peirson & Henley (1994). The (non)inclusion also depends on the period of interest for forecasting. During the winter in France for example, a decrease of 1^oC in the national temperature gave a 2100 MW estimated (by EDF internal model) increase in the national demand for winter 2008-09. The effect is especially large due to the relative importance of electric heating in France. Since accurate one-day ahead forecasts are available in France, it is common practice to include weather-based variables in forecasting models for this area. For a larger country such as Brazil with larger variation in temperature between regions, it can be hard to get a representative national temperature as well as accurate forecasts, see the discussion in Soares & Medeiros (2008). One of the aims in this thesis is to study weather influence on electricity load in France during a long period.

1.1.4 EDF model and new issues related to market liberalization

Although most of the results described in this dissertation can be applied to hourly electricity demand from other countries after some model adjustments as discussed earlier, we focus on French national electricity demand. Indeed, market liberalization is a recent event in France and the main electricity producer and retailer, Electricité de France (EDF), investigates new methods for electricity demand forecasting. Some of the past research results can be found in Ménage, Panciatici & Boury (1988), Martin (1999) or Bruhns, Deurveilher & Roy (2005).

Historically, the EDF customer portfolio corresponded to almost the whole of France: there was no competitive market but local providers (independent from EDF) were present. EDF had "just" to forecast national demand only. The national curve is therefore controlled and much expertise is available to adjust forecasts. In recent years, national

demand is split between different electricity providers, so that the EDF portfolio is subject to more variations than in the past. Market participants, EDF in particular, have to forecast electricity demand for their own portfolio, a subgroup of the national demand. Forecasting models in operation at EDF therefore need some adjustments. Different strategies can be considered:

- Direct model estimation of the complete portfolio demand signal (EDF-signal). The signal is therefore less stable than the national one, both in level and shape. Getting accurate data for the complete portfolio demand is itself an issue.
- Since forecasting accuracy for national demand is good, we can use this forecast and substract the estimated demand for non-EDF customers. This strategy is efficient when the consumption behaviour of lost customers is known and can be accurately evaluated and forecasted. This hypothesis becomes more and more unreliable with time.
- Aggregate load forecasts of some customer portfolios: each subgroup of customers is supposed to have a specific homogeneous consumption behaviour so that different model specifications are involved. The overall forecast is therefore the sum of all subgroup forecasts. Misiti, Misiti, Oppenheim & Poggi (2008) address this issue. Data availability is also an issue in this case.

Bruhns et al. (2005) give some details of the modelling strategy for French national demand in use at EDF. Total hourly electricity load on day t and hour (or half-hour) h is modelled with the following independent parts :

$$load_{h,t} = weatherindependentload_{h,t} + weatherdependentload_{h,t} + \varepsilon_{h,t} \tag{1.1}$$

where $weatherindependentload_{h,t}$ captures the trend and the different levels of seasonality (weekly, yearly) in a multiplicative way, $weatherdependentload_{h,t}$ captures weather-related influences (heating, cooling, cloud-cover) and $\varepsilon_{h,t}$ is an error term. A trend is included to model the gradual increase of total demand, the yearly pattern is captured using Fourier coefficients and the weekly pattern is modelled using day type definitions. The daily pattern is taken care of by building one equation per hour of the day. The influence of the temperature in the winter is modelled by a transformation of temperature and cloud-cover data into an indoor representative temperature. This transformation includes unknown parameters which need to be estimated. This involves non-linearities in the model. The influence of cooling is modelled in the same fashion although not including cloud-cover. Model (1.1) is used for mid-term forecasting. Short-term forecasts are obtained using an autoregressive model applied to model (1.1) errors with fixed weights, including lags as long as the last two weeks. The model also considers special

adjustments for bank holidays, related bridge days as well as days with special tariffs. Model (1.1) gives a highly satisfactory forecasting performance for national data but the price to pay is the large number of unknown parameters to estimate and the large number of years required for a proper estimation. Efficient constrained optimization is performed for fast estimation. Note that the model includes trends in the level but also for the heating coefficient for parameter extrapolation from one year to another in the forecasting process: these trends are difficult to set and require much expertise. It is expected that applying the current internal model directly to the EDF signal will give less accurate forecasts in the future due to the growing non-stability of the data.

To address the problem of forecasting the demand of a non-stable customer portfolio, mixing predictors is an effective strategy described in Goude (2008). A competitive strategy is to consider the use of adaptive models: we therefore discuss in this thesis state-space modelling including time-varying regression effects for French national demand. We prefer to build models on French national demand first because these data are well-known by internal forecasters at EDF so that empirical findings can be confirmed (or invalidated) while it is harder for the EDF-signal due to its recent set up (only 3 years of data are available). The adaptation of our modelling strategies for an efficient forecasting of the EDF-signal is left for further empirical investigations. The models described in the following chapters are inspired by model (1.1).

1.2 Introduction to state-space modelling

After describing the motivation of using adaptive models for short-term load forecasting, we present the state-space framework. Model specification, estimation and practical implementation are discussed as well as model diagnostics. The section ends with a short review of state-space models applications to electricity load.

1.2.1 Multivariate linear Gaussian state-space models

We consider the multivariate linear Gaussian state-space model written as follows:

$$\left\{ \begin{array}{rcll} y_t & = & Z_t\alpha_t + \varepsilon_t, & \varepsilon_t \sim N\left(0, H_t\right), \\ \alpha_{t+1} & = & T_t\alpha_t + R_t\eta_t, & \eta_t \sim N\left(0, Q_t\right), \end{array} \right. \quad t=1,\ldots,\mathrm{T}. \tag{1.2}$$

The first equation in (1.2) is the so-called observation equation where y_t is the $k \times 1$ observed variable to model and α_t is the state vector of dimension $m \times 1$, Z_t is the observation matrix and ε_t is a zero mean Gaussian noise with covariance matrix H_t. The

second equation in (1.2) is the transition equation where T_t is the transition matrix, η_t is a zero mean Gaussian noise, independent from ε_t, with covariance matrix Q_t, R_t is the error loading matrix. Textbooks such as Harvey (1989) and Durbin & Koopman (2001) give a complete description of the state-space methodology.

The Multivariate Linear Gaussian state-space model provides a unified framework for several classical statistical methods: SARIMA models of Box & Jenkins (1970), linear time-varying or constant regression models, exponential smoothing, cubic spline regression, etc. It also allows for non-stationary models. We provide some particular simple specifications that are used in the following chapters.

Local linear trend model:

$$\left\{\begin{array}{lcll} y_t & = & \mu_t + \varepsilon_t, & \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \\ \mu_{t+1} & = & \mu_t + \nu_t + \xi_t, & \xi_t \sim N(0, \sigma_\xi^2), \qquad t = 1, ..., T. \\ \nu_{t+1} & = & \nu_t + \zeta_t, & \zeta_t \sim N(0, \sigma_\zeta^2), \end{array}\right. \tag{1.3}$$

In model (1.3), the series of observations y_t is univariate and μ_t is the stochastic trend with slope term ν_t. ε_t, ξ_t, and ζ_t are zero mean Gaussian noises with respective standard deviations $\sigma_\varepsilon, \sigma_\xi$, and σ_ζ. Different particular cases fit in this specification: the classical deterministic trend when $\sigma_\xi = \sigma_\zeta = 0$. Taking only $\sigma_\zeta = 0$ gives the random-walk plus drift model and only $\sigma_\xi = 0$ gives the integrated random-walk model. Large values in σ_ξ or σ_ζ compared to σ_ε give an estimated trend that tends to follow the data (including the noise) while smaller values for σ_ξ or σ_ζ will give a smooth curve.

Stochastic regression model:

$$\left\{\begin{array}{lcll} y_t & = & x_t\beta_t + \varepsilon_t, & \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \\ \beta_{t+1} & = & \beta_t + \chi_t, & \eta_t \sim N(0, \sigma_\chi^2), \end{array}\right. \qquad t{=}1,\ldots,\mathrm{T}. \tag{1.4}$$

In model (1.4), the series of observations y_t is univariate and β_t is the stochastic regression coefficient associated to the explanatory variable x_t. ε_t and χ_t are zero mean Gaussian noises with respective standard-deviations σ_ε and σ_χ. This specification includes the constant parameter regression model for explanatory variables x_t when σ_χ is zero. Otherwise we obtain the time-varying regression model.

When model (1.2) is fully specified and there is no unknown parameter, the Kalman filtering and smoothing algorithms can be used for prediction and estimation of the unobserved state vector α_t. We detail unknown parameter estimation next.

Kalman filtering and smoothing equations when α_1 is known Denote by a_t the one-step ahead forecast of the state vector α_t based on past observed values $Y_{t-1} =$

$(y_1, \ldots, y_{t-1})$, $a_t = \mathbb{E}(\alpha_t|Y_{t-1})$, and $P_t = \mathrm{Var}(\alpha_t|Y_{t-1})$ the associated variance. The Kalman filtering equations compute recursively these values for $t = 1, \ldots, T$:

$$\begin{cases} v_t = y_t - Z_t a_t, & F_t = Z_t P_t Z_t' + H_t, \\ K_t = T_t P_t Z_t' F_t^{-1}, & L_t = T_t - K_t Z_t, \\ a_{t+1} = T_t a_t + K_t v_t, & P_{t+1} = T_t P_t L_t' + R_t Q_t R_t', \end{cases} \qquad t = 1, ..., T. \tag{1.5}$$

In equation (1.5), v_t is the one-step ahead forecast error for observation y_t and $F_t = \mathrm{Var}(y_t|Y_{t-1})$. Further, K_t is the Kalman gain matrix and L_t is an intermediate matrix. Estimates based on the full sample are deduced from the Kalman smoothing algorithm. Denote by $\hat{\alpha}_t$ the estimation of the state vector α_t based on all data and V_t the associated variance, $\hat{\alpha}_t = \mathbb{E}(\alpha_t|Y_T), V_t = \mathrm{Var}(\alpha_t|Y_T)$. The Kalman smoothing equations are based on the backward recursions for $t = 1, \ldots, T$:

$$\begin{cases} r_T = 0, & N_T = 0, \\ r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, & N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \\ \hat{\alpha}_t = a_t + P_t r_{t-1}, & V_t = P_t - P_t N_{t-1} P_t, \end{cases} \qquad t = 1, ..., T. \tag{1.6}$$

Initialization of the state vector α_1 An intricate part of the state-space model specification is the initial state value of the system α_1. In most cases it is unknown, and the usual specification is to suppose $\alpha_1 \sim N(a_1, P_1)$. The state vector α_t can be composed of constant, stationary and/or nonstationary components so that we write α_1 as $\alpha_1 = a + A\delta + R_0\eta_0$ where a is a known constant, δ is a diffuse (nonstationary) component and η_0 is a stationary component. We have $\eta_0 \sim N(0, Q_0)$ and $\delta \sim N(0, \kappa I_q), \kappa \longrightarrow +\infty$. We can therefore write $P_1 = \kappa P_\infty + P_*, P_\infty = AA', P_* = R_0 Q_0 R_0'$. De Jong (1991) and more recently, chapter 5 of Durbin & Koopman (2001) give theoretical developments of diffuse initial conditions for the Kalman filter and smoother. For non-degenerate models (i.e. containing enough information in the data to initialize the state vector), there is always a time $d, d << T$ where infinite components involving κ vanish and usual Kalman filtering and smoothing can be used.

Missing data A positive side of the state-space framework is its easy treatment of missing data. For the Kalman filter (1.5), the corresponding error term v_t is simply zero so that the filtered state vector estimate a_t remains unchanged. However, the associated variance F_t increases. The Kalman smoother allows to replace missing data with estimates based on the rest of the data. For example, parameter estimation in a SARIMA model with missing data is straightforward in the state-space framework.

1.2.2 Parameter estimation

In most cases, the state-space model (1.2) depends on unknown parameters in matrices T_t, Z_t or covariance matrices H_t, Q_t. These parameters (also called hyperparameters), put together in vector ψ, are usually estimated by maximizing the loglikelihood of the data. The loglikelihood function when α_1 is known is simply computed with one application of the Kalman filter using the prediction error decomposition:

$$\log L(y, \psi) = -\frac{Tk}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{T} \Big(\log |F_t| + v_t' F_t^{-1} v_t \Big) \tag{1.7}$$

The treatment of diffuse initial conditions requires minor modifications, see Ansley & Kohn (1985), De Jong (1991) or Koopman (1997).

Note that, by definition, maximizing the loglikelihood function is approximately equivalent to minimizing the Root Mean Squared Error (RMSE) of one-step ahead in-sample forecasts.

Two methods can be used to perform likelihood maximization:

- use a Scoring algorithm where the score vector is evaluated analytically or numerically, see the exact formulae and discussion in Koopman & Shephard (1992).
- use the well-known EM algorithm, detailed for state-space models in Shumway & Stoffer (1982) or Koopman (1993). It is used for parameters in H_t and Q_t. The EM algorithm is fast in its early steps and then converges slowly to the maximum likelihood estimates. This is why an intermediate strategy consists in using the EM algorithm as a first step to obtain good initial values for a Scoring algorithm in a second step.

If some elements in the vector of (hyper-)parameters ψ are subject to constraints, an appropriate transformation can be applied to ψ to perform unconstrained optimization. When estimating a constrained parameter such as an irregular term standard-deviation $\sigma > 0$, the so-called pile-up problem can occur. More precisely, it can happen that the parameter, which is supposed to be strictly positive, is estimated exactly zero. The pile-up issue is discussed in Shephard (1993). Stock & Watson (1998) suggest an alternative estimator of the standard-deviation parameter.

Estimation of constant regression coefficients We described earlier how to put constant regression effects in state-space form. There are therefore two ways to estimate an unknown regression coefficient:

- put the regression coefficients in the state vector α_t and estimate it recursively by means of the Kalman filter with diffuse initial condition for the coefficient. In this case, the marginal likelihood for the remaining parameters is computed, see Tunnicliffe Wilson (1989) for a discussion, or Francke (2006) and Francke, Koopman & de Vos (2008).
- put the regression coefficients in the vector of hyperparameters ψ and estimate it directly with the other parameters during the likelihood maximization step. In this case, the profile likelihood is maximized, as the regression parameters satisfy the standard first order conditions after optimization.

1.2.3 Model evaluation

If the model is well specified, then for the one step ahead forecast errors v_t, we have $v_t \sim N(0, F_t)$ and v_t and $v_{t'}, t \neq t'$ are independent. Two diagnostics are therefore used in the remaining of the dissertation:

- the empirical autocorrelation function (ACF) of the row by row in-sample one-step ahead standardized residuals $\hat{v}_{t,i}/\sqrt{(\hat{F}_t)_{i,i}}$ should exhibit no dynamic structure,
- the empirical distribution of the row by row in-sample one-step ahead standardized residuals $\hat{v}_{t,i}/\sqrt{(\hat{F}_t)_{i,i}}$ should be approximately Gaussian with empirical mean close to 0 and variance close to 1.

Another model evaluation tool is the post-sample forecasting accuracy. Computing $t+h, h \geq 1, t > T$ forecasts is straightforward in the state-space framework. It only requires the Kalman filter based on maximum likelihood estimates of unknown parameters and the estimate of the final state vector $\hat{\alpha_T}$ treating observations $t+1, \ldots, t+h$ as missing data. Since the likelihood maximization step minimizes in-sample residuals RMSE, it is expected that the model also gives satisfactory RMSE for the post-sample. Note that another recurrent accuracy measure in load forecasting used in the following chapters is the MAPE, introduced in the literature review in §1.1.3, which is not optimized by the likelihood maximization but is less affected by outliers.

1.2.4 Practical implementation

The thesis investigates three modelling strategies for hourly electricity loads in France. All computations are done in `Ox`, a matrix object-oriented language, see Doornik (2006), and econometrics software `OxMetrics`. Some built-in maximization methods are available in particular the BFGS (Broyden-Fletcher-Goldfarb-Shannon) algorithm for func-

tion maximization using first derivatives (calculated numerically or analytically), see e.g. Fletcher (1987) for details. For state-space models specification, estimation, signal extraction, forecasting and diagnostic functions `SsfPack` packages provides ready-to-use routines, see Koopman, Shephard & Doornik (1999) and Koopman, Shephard & Doornik (2008).

1.2.5 Kalman filtering and Electricity load modelling

Some papers suggest state-space modelling for short-term electricity demand forecasting. However models are often limited to stationary ARIMA components and constant regression effects. Al-Hamadi & Soliman (2004) built a periodic hourly independent ARMAX model with the temperature as explanatory variable for Canadian electricity load data. ARMAX coefficients were constant and estimated for each hour using the Kalman filter on the state-space model. They built separate models for weekdays and weekends.

State-space modelling for load forecasting is not new for EDF. Martin (1999) suggested a seasonal state-space model for load measured every 5 minutes. Nested Kalman filters were used to reduce the state vector dimension. The model is equivalent to a SARIMA specification and applied to the load corrected for weather effects. Intervention variables were included in the model as well.

Harvey & Koopman (1993) suggested a structural univariate time series model for hourly electricity loads involving time-varying splines to model the intraweekly pattern. The temperature effect was modelled by a constant spline regression and the model included a time-varying trend and a deterministic cycle. The model was used in the same forecasting competition as Ramanathan et al. (1997), the latter being slightly more accurate.

Zheng, Girgis & Makram (2000) combined the wavelet transform and state-space modelling by including wavelet coefficients in the state vector and modelled them by random-walks. The observation equation matrix contained the orthogonal matrix of the wavelet transform to reconstruct the load. They applied their method to a local utility data. Weather dependence was included, however model parameters were set empirically. Gastaldi, Lamedica, Nardecchia & Prudenzi (2004) also used local (municipal) data but built a univariate time series model for hourly loads without exogenous effects, restricted themselves to weekdays and modelled hourly load variations to forecast the future load. Pedregal & Young (2008) adopted the unobserved component model framework with nonlinear exogenous variables effects by means of radial basis functions. They used a backfitting algorithm for parameter estimation and the Kalman filtering and smoothing

algorithms for signal extraction. The method was applied to very local electricity demand data, modelled as a univariate time series. They considered the long-term trend and the annual cycle in the same model component since the two are difficult to estimate separately and this did not appear to affect forecasting accuracy.

1.3 Overview of the thesis

1.3.1 Chapters description

The dissertation is a compilation of three papers dedicated to state-space model specifications for French national hourly electricity loads. Data are considered as daily vectors of hourly loads. Throughout the thesis we present complementary model specifications within the Multivariate Linear Gaussian state-space framework and show that the different methodologies are effective for the modelling of French national hourly load. The methodologies are flexible enough to capture highly as well as smoothly time-varying features and help the understanding of the different component dynamics in electricity demand.

Chapter 2 details a first modelling approach. A periodic dynamic linear regression model is specified for the daily vector of hourly loads and is estimated on a long dataset (10 years). The model includes stochastic trends, weekly and yearly seasonal components as well as stochastic regression effects for weather-based exogenous variables (the regression coefficient for the cloud-cover is constant). Constant regression effects for special days are also considered. The multivariate model specification involves non-diagonal covariance matrices for the different stochastic components. This choice allows the estimation of multiple 24×24 possibly full rank covariance matrices which is an intricate problem. Considering the vast amount of data, empirical estimation is performed for a selection of hours only. In-sample signal extraction results for the different components of electricity demand are detailed for the interesting couple (9 AM, 12PM). The choice in the couple of hours under study is arbitrary but this couple is of particular interest for EDF. It is more interesting to focus on peak hours such as 9 AM when demand is high (and therefore producing costs as well). Day hours are more interesting than night hours because they are more difficult to forecast. The model is able to capture long-term smooth variations for the cooling effect and also for Monday/Friday effects, but is also able to capture fast and large variations such as the August effect. Another interesting feature is the non-linear evolution of the heating regression coefficient within the winter period. Forecasting accuracy is evaluated, for the (9 AM,12 PM) model and

more generally for each couple of successive hours: post-sample forecasts are computed from one day up to one week ahead and compared with benchmark models including the univariate dynamic regression model that is based on the multivariate one. Both the univariate and multivariate specifications give good forecasting accuracy, respectively 1.41% and 1.40% for non-special days. Weekly forecasts however suggest possible improvement in the transition equations specification for the stochastic components. The chapter also discusses the impact of using forecast or actual weather data for model evaluation. Analysing estimated correlations, empirical results suggest that some restrictions can be imposed in the dimension of the stochastic components. For practical reasons, we could not consider more than two hours in the same model. The following chapters present complementary specifications that allow to consider more hours together. The chapter has been published in the International Journal of Forecasting, as Dordonnat, Koopman, Ooms, Collet & Dessertaine (2008)[2] .

Chapter 3 also describes a periodic dynamic linear regression model and more specifically the specification of dynamic factors to reduce dimension in the stochastic components of electricity load. The idea is that a specific stochastic component is the same for a subset of hours up to a linear transformation. The general dynamic factor regression model is presented as well as particular specifications of interest. The model specification is first motivated with a simulation study that compares the factor model with the independent modelling of each hourly element of vector y_t from two perspectives: parameter estimation and signal extraction. Then a model is suggested for French national hourly loads (special days are excluded from the empirical study), where the stochastic components are similar to the ones of chapter 2 but restricted to a block diagonal specification for the factor loading matrices. It follows that the model is estimated independently by group of three hours. In-sample signal extraction for all submodels is discussed, intraday evolution of the dynamic components in particular. The cooling effect is harder to detect with this model specification. The model still detects strong variations in the yearly pattern. Comparison is based on the univariate model specification at 9 AM. Most results are consistent, some components are however more varying with the univariate specification. The model is overall satisfactory in terms of one-day ahead forecasting accuracy, slightly inferior to the univariate model specification. A shorter version of this chapter is forthcoming in the proceedings of the IEEE Power Engineering Society (PES) General Meeting, 2009, as Dordonnat, Koopman & Ooms (2009).

We investigate in chapter 4 the combination of the dynamic factor modelling approach of chapter 3 with piecewise regression splines. The intradaily pattern is modelled by the

[2]DOI:10.016/j.ijforecast.2008.08.010

spline curve with number of knots and their positions (abscissa) at certain hours of the day imposed a priori. The load coordinates of the knots themselves are modelled using periodic dynamic regression models. This leads to a dynamic factor regression model where factor loadings are determined a priori. The stochastic components for the electricity load are similar to the model specifications of chapters 2 and 3. Hourly exogenous variables however require preliminary transformation to be adapted to the smoothing spline. The estimation step involves a transformation of the load data to accelerate computations. The dataset used for empirical results is an extension of the previous chapters' datasets. Univariate models and dynamic factor models from chapter 3 are also applied to the same data to compare results. The suggested model is far more parsimonious in the number of parameters to estimate than our benchmark models. Our new model gives satisfactory results for most of the daily variables effects but somewhat disappointing results for the heating effect. The estimated daily effects (weekly and yearly pattern) are consistent with internal model analysis from EDF. Forecasting accuracy is relatively poor but this has not been the primary aim of this chapter. We expect to improve the model with further empirical investigations as suggested in the chapter conclusion.

The concluding chapter 5 summarizes the main modelling aspects and empirical results of the dissertation and provides suggestions for further research.

1.3.2 Empirical applications data and sources

The main data used in this thesis correspond to French national hourly electricity demand, measured in MegaWatts (MW), as illustrated in Figures 1.1 and 1.2 above. Data can be downloaded from the French Transport System Operator (RTE, Réseau de transport d'électricité) website [3]. Calendar information is also used to identify bank holidays, bridge days, Christmas/New Year period (defined from December 23^{rd} until January 3^{rd} in the dissertation), daylight saving days (switch between summer and winter time), and PDW days [4]. Finally, weather data are also considered : these data are confidential, provided by Meteo France. We use temperature data, in Celsius degrees (^oC), actual values as well as forecasts. We also use observed cloud-cover data: this variable has no

[3] `http://www.rte-france.com/htm/an/vie/vie_stats.jsp`

[4] PDW (Peak Day Withdrawal or EJP - Effacement Jour de Pointe - in French) days correspond to winter days when national demand need to be reduced in large proportions: it can be the consequence of very cold temperature or electricity production problem. In this case, large electricity consumers with adequate contract have to reduce their demand or face important financial costs. PDW days can only be weekdays and only occur between November and March. The number of PDW days for each year is limited to 22; however rules changed recently: PWD are set locally (France is then divided in 4 regions) so that 88 PDW days by year can theoretically occur.

real unit (the sky is divided in 8 cells and cloud-cover is defined as the number of cloudy cells), it is based on human/machine observation and integer-valued - 0 means the sky is clear, 8 means the sky is very cloudy. Effective weather data used for empirical models are weighted averages of local temperatures and cloud-covers from 26 meteorological stations; weights have been determined internally at EDF. We do not use other weather variables such as wind speed or humidity which are sometimes considered for short-term electricity load forecasting, see e.g. Contaxi, Delkis, Kavatza & Vournas (2006) who defined equivalent temperature by taking account of high levels of relative humidity for Cyprus data or Fan & Mc Donald (1993) who also incorporated wind speed in equivalent temperature. An approach for French data is proposed by Laluque (2007).

In chapter 2, the period of interest is September 1^{st}, 1995 until August 31^{st}, 2004. Chapter 3 results are restricted to the period January 1^{st}, 1997 until August 31^{st}, 2004. Chapter 4 uses an extended data set: from January 1^{st}, 1997 until August 31^{st}, 2007. In all applications, the dataset is split in two parts: the first years are used for model estimation and the last year of data is used for post-sample forecasting evaluation.

Missing data Electricity and weather data are complete and all observations are valid. However, in the following chapters some special days and periods are explicitly considered as missing:

- The modelling strategy for bank holidays, bridge days, Christmas/New Year period and daylight saving days in chapter 2 is very basic and leads to inaccurate short-term forecasts. Improving these forecasts requires much expertise because these events strongly alter the daily load curve, both in shape and level. The effect depends on which day of the week (Christmas) or the year (Easter) the special day of interest occurs so that only few data can be used for each calendar configuration. In this dissertation, we are more interested in general electricity consumption behaviour and their long-term or intra-yearly evolution, we therefore exclude these days from the data in chapters 3 and 4 : they are treated as missing data.
- PDW days also strongly affect the daily load curve and require a specific modelling strategy. In the whole dissertation, these days are considered as missing data for simplicity. Chapter 2 discusses this topic in more detail.

The modelling and forecasting for all these special days is important to provide an operational model (PDW days represents e.g. 7.5% of the data), however it would not help the comprehension of the methodology. This practical aspect can be considered for future improvement of the models.

Chapter 2

An Hourly Periodic State Space Model for Modelling French National Electricity Load

Abstract We present a model for hourly electricity load forecasting based on stochastically time-varying processes that are designed to account for changes in customer behaviour and in utility production efficiencies. The model is periodic: it consists of different equations and different parameters for each hour of the day. Dependence between the equations is introduced by covariances between disturbances that drive the time-varying processes. The equations are estimated simultaneously. Our model consists of components that represent trends, seasons at different levels (yearly, weekly, daily, special days and holidays), short-term dynamics and weather regression effects including nonlinear functions for heating effects. The implementation of our forecasting procedure relies on the multivariate linear Gaussian state space framework and is applied to national French hourly electricity load. The analysis focuses on two hours, 9 AM and 12 PM, but forecasting results are presented for all twenty-four hours. Given the time series length of nine years of hourly observations, many features of our model can be estimated readily including yearly patterns and their time-varying nature. The empirical analysis involves an out-of-sample forecasting assessment up to seven days ahead. The one-day ahead forecasts from fourty-eight bivariate models are compared with twenty-four univariate models, one for each hour of the day. We find that the implied forecasting function strongly depends on the hour of the day. This chapter has been published in the International Journal of Forecasting, as Dordonnat et al. (2008)[1] .

[1]DOI:10.016/j.ijforecast.2008.08.010

2.1 Introduction

There is a need for accurate forecasts of the electricity load in both the short and long term. Short-term forecasting is important because the national grid requires a balance between the electricity produced and consumed at any moment in the day. Long-term forecasting is relevant for the planning of new electricity utilities, and inaccurate forecasts have important financial costs. This chapter aims to develop an effective new method for the short-term forecasting of hourly electricity loads.

During the past years, many papers have been dedicated to methods and models for hourly load forecasting. Contributions can be distinguished between statistical models and exponential smoothing methods, between univariate models and models with explanatory variables, between linear models and nonlinear models. Earlier papers have developed both single-equation models and multiple-equation models with different equations for the different hours of the day. Both independent multiple-equation models and correlated multiple-equation models have been specified. The time dependence of hourly loads has been captured in both observation-driven VARIMA-type models, and parameter-driven models with unobserved components. In this chapter we develop a forecasting model based on an interpretable decomposition of electricity loads in a trend, time-varying seasonal effects, calendar effects and weather dependent effects.

Our model is inspired by Ramanathan et al. (1997) who built an extensive multiple regression model with separate forecasting equations for each hour of the day. Their observation-driven model included calendar and weather effects and outperformed a wide range of alternative models in a forecasting competition. Taking part in the same competition, Harvey & Koopman (1993) developed an unobserved components model with time-varying splines to capture the evolution of intradaily seasonal patterns of hourly electricity loads, thereby integrating the equations for the different hours of the day. In a model for the New South Wales electricity load Cottet & Smith (2003) also used a multi-equation approach to capture the intradaily pattern, and developed long and short-term forecast models within a Bayesian framework; however, they assumed a diagonal vector autoregressive structure for the error terms. In this chapter we follow Smith & Kohn (2002) in allowing cross-correlation between the stochastic terms of the equations for the different hours of the day.

Our model for the French load includes all of the well-known features in electricity consumption, see e.g., Bunn & Farmer (1985), who studied different levels of seasonality (yearly, weekly, daily), and the effects of calendar events and weather dependence; and Cancelo & Espasa (1996), who built a single equation model for daily electricity loads,

thoroughly investigating the effects of special days and the relationship between the electricity load and temperature. The flexibility of their approach is illustrated by Cancelo, Espasa & Grafe (2008). Bruhns et al. (2005) gave a detailed description of a non-linear forecasting model of French load in use at Electricite de France (EDF), which allowed for different levels of seasonality and weather dependence. In this chapter, we present a different multiple-equation linear time-varying regression model for French national hourly electricity load, with one equation for each hour, like the approach of Ramanathan et al. (1997), and more recently, Soares & Souza (2006). We do not include periodic seasonal ARIMA components as they are difficult to interpret from an economic point of view. Instead, like Harvey & Koopman (1993) we capture the dynamics using a time-varying parameter regression, in order to understand the possible causes of changing trends and seasonal patterns. Pedregal & Young (2006) find periodic parameter changes at different frequencies in their analysis of twelve weeks of four-hourly load data in a dynamic harmonic regression model. Like Young, Pedregal & Tych (1999), they were unable to identify yearly movements. In addition to changes according to the hour of the day we discover a yearly pattern in the effect of temperature, which we partly model as a nonlinear heating effect.

We do not claim that univariate methods deliver bad forecasts. Structural univariate modelling of hourly demand has been suggested by Martin-Rodriguez & Caceres-Hernandez (2005), who proposed the use of unobserved component models and splines to capture the different levels of seasonality in the data. Other authors have preferred to build a model without weather variables. They argue that the availability and accuracy of weather data forecasts can be problematic, see the interesting discussion in Soares & Medeiros (2008). Taylor et al. (2006) compared the forecasting performance of a wide range of univariate time series methods for intradaily load forecasting. Taylor & McSharry (2007) studied the effectiveness of these methods for forecasting hourly and half-hourly loads in 10 European countries in the period May-September 2005 and found the performance of recent univariate methods quite promising. Taylor & Buizza (2003) exploit different scenarios in temperature forecasting to estimate its effect on load forecasting uncertainty. Following Taylor & McSharry (2007) we use a simple univariate weekly random walk model as a benchmark in our forecast evaluation.

The first aim of our study is to examine the evolution of the effect of explanatory variables over a long period via the time-varying structure of our model. This evolution may be related to the gradual market penetration and slowly changing efficiency of electricity utilities, for example to the heating and cooling effects. These adjustments are also determined by changes in users' behaviour, for example the Friday afternoon

effect. Such changes are not independent across the different hours of the day, so we study the different hourly loads in one joint model, allowing for cross-equation correlations in the innovations. The second modelling aim is to provide accurate short-term forecasts, from one day to one week ahead. The interpretation of the time-varying effects helps us in understanding possible forecast inaccuracies.

We show that our model fits in the multivariate linear Gaussian state space models framework. This implies that we can use Kalman filtering and associated algorithms to estimate the different components of the electricity load and to do short-term forecasting. The model parameters are estimated by maximum likelihood. The interpretation of the results is consistent with expert analysis from EDF, the French national electricity producer and provider. Except for some special thresholds in the temperature effect we confine ourselves to linear methods, in contrast to Engle, Granger, Rice & Weiss (1986), Liu, Chen, Liu & Harris (2006), Cottet & Smith (2003) and Hippert et al. (2005), who consider semiparametric methods and artificial neural networks for modelling meteorological effects and seasonal patterns.

The plan of the remainder of this chapter is as follows. Section 2.2 describes the dataset, Section 2.3 details the building of the model, describes the construction of the explanatory variables and relates the model to the linear state space framework. Section 2.4 presents the estimation results and interprets the various time-varying patterns. Section 2.5 discusses the absolute and relative forecasting performance. Section 2.6 concludes.

2.2 Description of French national hourly electricity loads

The dataset used in this study concerns French national hourly electricity consumption from September 1, 1995 until August 31, 2004. This hourly time series is for nine years and consists of $3,288$ daily (or $78,912$) hourly observations. The dataset is compiled by Electricité de France (EDF) and is complete; that is, no missing observations are present. However, some days are intentionally considered as missing and excluded from the analysis in this study. On these days, which are known at EDF as EJP (Effacement Jour de Pointe: Peak Day Withdrawal) days, the load supply is subject to special tariffs. These financial incentives are introduced to cut heavy consumption. The EJP tariffs are activated when high levels of consumption and/or when problems with electricity supply (production) are expected to occur. They can only be set in place between November

and March, and are for working days only. Clearly, on these days, the daily load and the hourly load curve are severely affected, and standard models will overestimate electricity demand. Special treatments for the forecasting of the EJP days are outside the scope of this chapter. The number of EJP days in our dataset is 249, that is 7.5% of the total number of days.

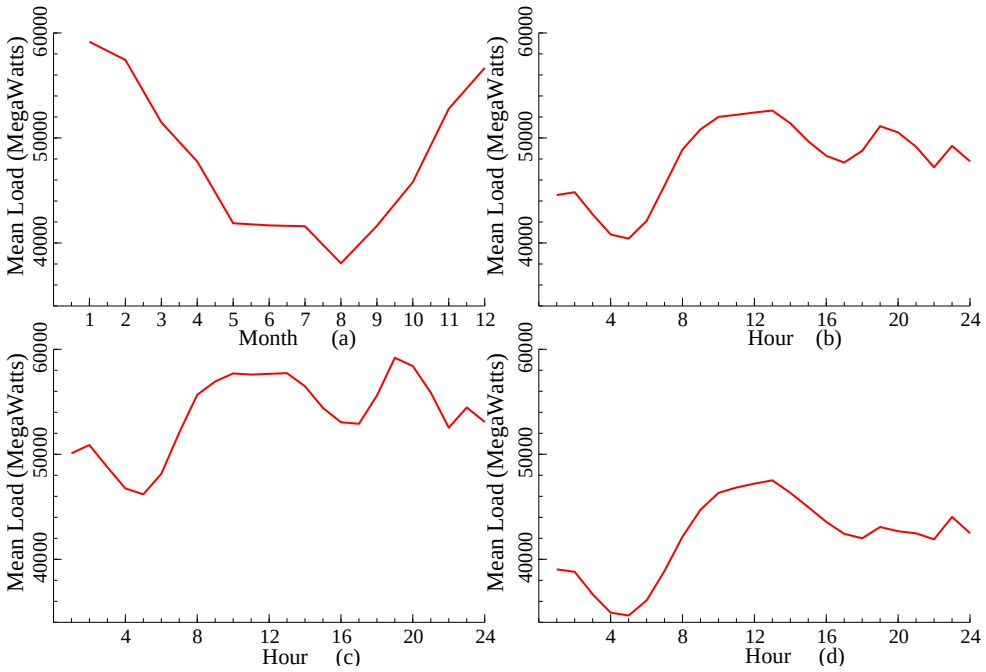

Figure 2.1: Data description of French national electricity loads from September 1, 1995 to August 31, 2004: (a) monthly averages; (b) hourly averages; (c) hourly averages for the winter months (October – March); (d) hourly averages for the summer months (April – September).

Different averages of the French electricity consumption are presented in Figure 2.1. Panel (a) shows the monthly averages for the complete sample and shows that the consumption is highest in January, and lowest in the holiday month of August. The other panels present the hourly averages for (b) the complete sample, (c) the winter months, and (d) the summer months. In all cases, the lowest electricity consumption is observed at 5 AM, while the consumption is highest at 1 PM except in the winter months when the consumption is highest at 7 PM. Apart from the levels of consumption, the intra-daily winter and summer electricity load curves are quite similar, except in the early evening hours of 6 – 9 PM, when consumption increases during the winter period only.

In addition to calendar information for holidays and other special days, the dataset also includes hourly temperature, cloud cover measures, and one-day ahead forecasts of the hourly temperature. The source of these data is Meteo France, and they provide

the data for different regions in France. Measures of cloud cover are based on human observations, and forecasts of cloud cover are not provided. EDF weights the temperature and cloud cover data for the different regions to construct a national average for the hourly average temperatures, their one-day forecasts, and cloud cover. The use of weather variables in our forecasting model is regarded as crucial at EDF since much heating is generated by electricity in France. Approximately 28% of the private homes in France have electric heating.

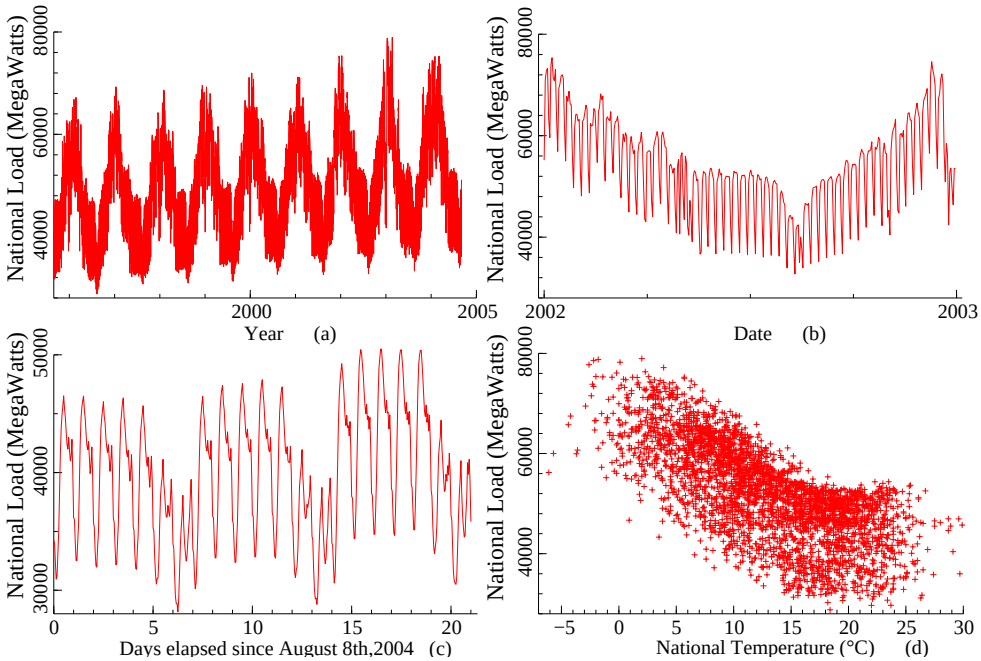

Figure 2.2: Data description: (a) French national daily electricity loads from September 1, 1995 to August 31, 2004, at 9 AM; (b) Electricity loads at 9 AM in 2002 (including special tariff EJP days); (c) Hourly electricity loads in the three weeks after August 8, 2004; (d) Daily electricity loads versus the national average temperature from September 1, 1995 to August 31, 2004, at 9 AM.

Figure 2.2 provides some further graphical insights into our dataset. Panel (a) presents the daily French national electricity consumption at 9 AM for the sample period of our dataset. The yearly seasonal cycle is clearly visible, and a positive trend is also detectable from the data. These features also appear to various extents at the other hours of the day. Panel (b) presents the load at hour 9 AM for the year 2002. In this graph the weekly seasonality of the electricity load becomes apparent, as do the effects of the summer holiday in August and the special tariff EJP days. The last three weeks of the hourly loads in our dataset, that is from Monday, August 8, 2004 until Sunday, August 29, 2004, are presented in panel (c). This shows the magnitude of the weekly and daily load curves.

The decreases in the load during Saturday, Sunday and the night hours of all weekdays in the week are all distinct from each other and from the working day hours generally. Panel (d) presents the well-known non-linear relationship between electricity load and average national temperature at 9 AM. The break in the regression curve appears to be at around 15 ^oC.

2.3 Model specification and parameter estimation

Let $y_{t,i}$ be the electricity load at day t and hour i, measured in megawatts. The basic model for the electricity loads that we consider in our study is a seemingly unrelated regression equations (SURE) model where each equation represents a particular hour i. The hourly regression equation is given by

$$y_{t,i} = f_{t,i} + X'_{t,i}\beta_{t,i} + W'_t\gamma_{t,i} + \varepsilon_{t,i} \qquad \varepsilon_{t,i} \sim NID(0, \sigma^2_{\varepsilon,i}), \tag{2.1}$$

where $f_{t,i}$ is the trend component, $X_{t,i}$ is a vector of explanatory variables that change with day t and hour i, whereas W_t is a vector of explanatory variables that only change with day t, for $t = 1, \ldots, n$ and $i = 1, \ldots, k$, where k is typically equal to 24 in the case of hourly data. Examples of variables in $X_{t,i}$ are temperature and cloud cover, since these variables change with the hour. Examples of variables in W_t are dummies for day-type and holidays, since these variables do not change with the hour. We note that (2.1) can also be considered for a subset of hourly data, so that k can take a value of 2 when two specific hours are considered, for example. The disturbance (or irregular) $\varepsilon_{t,i}$ is a random term with mean zero and variance σ^2_i that can be different for different hours. Irregulars of different hours in the same day can also be correlated with each other, that is $E(\varepsilon_{t,i}\varepsilon_{t,j}) \neq 0$ for $i \neq j = 1, \ldots, k$. Irregulars of different days are not correlated, that is $E(\varepsilon_{t,i}\varepsilon_{s,j}) = 0$ for $t \neq s = 1, \ldots, n$ and $i, j = 1, \ldots, k$.

Under the following conditions, the SURE system is standard, and the estimation of the unknown regression coefficients can take place using generalised least-squares methods:

(a) the trend component $f_{t,i}$ is a deterministic function of time; for example, $f_{t,i} = a_i + b_i t$ with unknown and fixed regression coefficients a_i and b_i for $i = 1, \ldots, k$;

(b) the regression coefficients in vectors $\beta_{t,i}$ and $\gamma_{t,i}$ are unknown, fixed and the same for each day, that is $\beta_{t,i} = \beta_i$ and $\gamma_{t,i} = \gamma_i$, for $i = 1, \ldots, k$;

Since the regression coefficients are different for different hours (equations), we can refer to the SURE system as a periodic model (periodic in hours).

One focus of our study is the variation in the regression coefficients over the days. This time variation can be modelled explicitly. For example, we can specify the trend component $f_{t,i}$ as a stochastic function of time (the details are given below). The regression coefficients in the vectors $\beta_{t,i}$ and $\gamma_{t,i}$ of (2.1) become time-varying when we specify these as random walk coefficients. In the remainder of this section we will discuss the details of the model that we adopt in our empirical study.

2.3.1 Stochastic trend component

The trend component $f_{t,i}$ represents long-term changes in electricity consumption. A flexible stochastic specification of a time-varying trend component is given by the local linear trend model

$$\begin{cases} f_{t+1,i} &= f_{t,i} + g_{t,i} + v_{t,i}, & v_{t,i} \sim NID(0, \sigma^2_{v,i}), & i = 1, \ldots, k, \\ g_{t+1,i} &= g_{t,i} + w_{t,i}, & w_{t,i} \sim NID(0, \sigma^2_{w,i}), & t = 1, \ldots, n, \end{cases} \tag{2.2}$$

where $v_{t,i}$ and $w_{t,i}$ are mutually and serially uncorrelated Gaussian noise terms with mean zero and variances $\sigma^2_{v,i}$ and $\sigma^2_{w,i}$, respectively, for $i = 1, \ldots, k$. The disturbances $v_{t,i}$ and $v_{s,j}$ can only be correlated for $i \neq j = 1, \ldots, k$ and $t = s = 1, \ldots, n$. This correlation is also allowed for $w_{t,i}$ and $w_{s,j}$. Special cases of the local linear trend model (2.2) include the random walk (with $\sigma^2_{w,i} = 0$ and $g_{1,i} = 0$), the random walk with fixed drift (with $\sigma^2_{w,i} = 0$ and $g_{1,i} \neq 0$), the integrated random walk (with $\sigma^2_{v,i} = 0$) and the linear fixed trend (with $\sigma^2_{v,i} = 0$ and $\sigma^2_{w,i} = 0$). The local linear trend model has $y_{t,i} = f_{t,i} + \varepsilon_{t,i}$ but this specification can be extended with other stochastic components for stationary (cyclical) processes and time-varying seasonal components. Such models are referred to as structural time series models or unobserved components time series models and are discussed at length in Harvey (1989). From this textbook treatment, we can, for example, learn that the forecasting function of the local linear trend model (2.2) is the well-known non-seasonal Holt-Winters method. The discount coefficients of this forecasting scheme are determined by the variances of the local linear trend model, $\sigma^2_{v,i}$ and $\sigma^2_{w,i}$.

2.3.2 Fixed and time-varying regressions for hourly explanatory variables

The hourly explanatory variables in $X_{t,i}$ concern weather variables that are based on temperature and cloud cover. In the model used in the empirical study below, we include four $X_{t,i}$ variables, of which three are related to temperature and one is related to cloud

cover. The three constructed temperature variables are designed to approximate the non-linear relationship between electricity load and temperature into a linear relationship. Denote the national average temperature in ^oC at day t and hour i as $T_{t,i}$. The first three variables are based on $T_{t,i}$ and a smoothed temperature variable $T^{smo}_{t,i}$. The smoothed temperature is computed recursively by an exponentially weighted moving average of temperature $T_{t,i}$ of previous hours, that is

$$\begin{array}{rcll} T^{smo}_{t,i+1} & = & \kappa T^{smo}_{t,i} + (1-\kappa) T_{t,i+1}, & i = 1, \ldots, k-1, \\ T^{smo}_{t+1,1} & = & \kappa T^{smo}_{t,k-1} + (1-\kappa) T_{t+1,1}, & i = k, \end{array} \tag{2.3}$$

with κ being typically close to 1, for example, 0.98. The smoothed temperature $T^{smo}_{t,i}$ is designed to take into account the physical inertia of buildings, for example. The first three variables in $X_{t,i}$ are constructed by

$$\left\{ \begin{array}{rcll} X^1_{t,i} & = & \max(0, 15 - T_{t,i}), & \\ X^2_{t,i} & = & \max(0, 15 - T^{smo}_{t,i}), & i = 1, \ldots, k, \;\; t = 1, \ldots, n, \\ X^3_{t,i} & = & \max(0, T^{smo}_{t,i} - 18), & \end{array} \right. \tag{2.4}$$

where we refer to $X^1_{t,i}$ as the heating-degrees variable, $X^2_{t,i}$ as the smoothed-heating-degrees variable and $X^3_{t,i}$ as the smoothed-cooling-degrees variable. The threshold temperatures for heating (15 ^{o}C) and cooling (18 ^{o}C) have been fixed at values determined internally at EDF. The last weather variable used in our model is the cloud cover variable $X^4_{t,i}$ that represents national cloud cover.

The vector of hourly explanatory variables is therefore

$$X_{t,i} = \left(\begin{array}{cccc} X^1_{t,i} & X^2_{t,i} & X^3_{t,i} & X^4_{t,i} \end{array} \right)'.$$

The regression coefficients that determine the total hourly weather effect are collected in the vector of unknown coefficients

$$\beta_{t,i} = \left(\begin{array}{cccc} \beta^1_{t,i} & \beta^2_{t,i} & \beta^3_{t,i} & \beta^4_{t,i} \end{array} \right)'.$$

The evolution of these regression coefficients over time is a novelty in the modelling of hourly electricity loads. The time-varying coefficients are modelled by

$$\left\{ \begin{array}{rcll} \beta^j_{t,i} & = & \beta^{*j}_{t,i} + \lambda^j_i X^j_{t-1,i}, & j = 1, 2, \\ \beta^3_{t,i} & = & \beta^{*3}_{t,i}, & \\ \beta^{*j}_{t+1,i} & = & \beta^{*j}_{t,i} + u^j_{t,i}, & j = 1, 2, 3, \end{array} \right. \tag{2.5}$$

where the disturbance $u^j_{t,i}$ is distributed as NID $(0, \sigma^2_{u^j,i})$, is serially uncorrelated, and can only be correlated with $u^j_{t,m}$, for $t = 1, \ldots, n$, $i \neq m = 1, \ldots, k$ and $j = 1, 2, 3$. The coefficient $\beta^4_{t,i} = \beta^4_i$ for cloud cover is constant (it does not vary over the days t),

for $i = 1, \ldots, k$, because there is no clear reason to expect the effect of cloud cover to change over time. The fixed and unknown regression coefficient λ_i^j determines the dependence of the heating regression coefficient on the temperature of the previous day at the same hour, that is $X_{t-1,i}^j$ for $j = 1, 2$ and $i = 1, \ldots, k$. In this way we introduce a mild nonlinear temperature effect into the model since $X_{t,i}^j \beta_{t,i}^j = X_{t,i}^j \beta_{t,i}^{*j} + \lambda_i^j X_{t,i}^j X_{t-1,i}^j$. In case temperature does not change much between days, we have $X_{t,i}^j X_{t-1,i}^j \approx (X_{t,i}^j)^2$. Furthermore, we introduce a yearly periodic dependence in the model. Since a time series of average temperature has a strong yearly cycle, the coefficient $\beta_{t,i}^j$ also changes with the yearly seasons of winter and summer temperatures, $j = 1, 2$. This seasonal dependence of coefficients is also referred to as periodic. The model is periodic within the day (different coefficients for different hours) but also within the year (some coefficients depend on a yearly cycle via the temperature variable). Finally, the regression coefficient of the cooling effect is time-varying and modelled by a random walk process.

2.3.3 Fixed and time-varying regressions for daily calendar variables

The vector of explanatory variables that only change by day (not by hour) is denoted by W_t, and is mainly concerned with the measurement of yearly, weekly and daily seasonal effects. With respect to the yearly seasonal effect in the electricity load that is not captured by the temperature effects in $X_{t,i}$, we consider the following Fourier terms as explanatory variables:

$$a_{s,t} = \cos\left(\tau_t \frac{2\pi s}{365.25}\right), \quad b_{s,t} = \sin\left(\tau_t \frac{2\pi s}{365.25}\right), \quad s = 1, \ldots, 4, \tag{2.6}$$

where τ_t is the number of days elapsed since the 1^{st} of January in the year in which day t falls for $t = 1, \ldots, n$. Moreover, we make a distinction between weekdays on the one side and weekends/holidays on the other side. For this purpose, we specify

$$a_{s,t}^{WD}, b_{s,t}^{WD} = \begin{cases} a_{s,t}, b_{s,t}, & \text{if day } t \text{ is a weekday;} \\ 0, & \text{if day } t \text{ is a weekend;} \end{cases} \tag{2.7}$$

and

$$a_{s,t}^{WE}, b_{s,t}^{WE} = \begin{cases} 0, & \text{if day } t \text{ is a weekday;} \\ a_{s,t}, b_{s,t}, & \text{if day } t \text{ is a weekend.} \end{cases} \tag{2.8}$$

As a result, the yearly cycle for electricity load is modelled by 4 Fourier series that requires 8 coefficients (for the cosine and sine parts) for the weekday yearly cycle and another 8 coefficients for the weekend yearly cycle. The variables $a_{s,t}^{WD}$, $b_{s,t}^{WD}$, $a_{s,t}^{WE}$ and $b_{s,t}^{WE}$ for $s = 1, 2, 3, 4$ are the first 16 explanatory variables in the vector W_t.

The typical weekdays of Tuesday, Wednesday and Thursdays (if not holidays) are taken as the default day effect in the model, and, obviously, no explanatory variable is introduced for this default day, to avoid multicollinearity problems. For the weekly seasonal effect and other calendar effects, we introduce a range of dummy variables that correspond to different day types, and are based on the operational practices at EDF, see Table 2.1.

Table 2.1: Daily explanatory variables

$W_t^{1,\dots,16}$	Fourier series for weekdays and weekends;
	Dummy variables for day types:
W_t^{17}	Mondays (if not a holiday or bridge day);
W_t^{18}	Fridays (if not a holiday or bridge day);
W_t^{19}	Saturdays;
W_t^{20}	Sundays;
W_t^{21}	Holiday (Easter Monday, Ascension Day, Whit Monday, May 1^{st}, May 8^{th}, July 14^{th}, August 15^{th}, November 1^{st}, November 11^{th}, if not a Saturday or Sunday);
W_t^{22}	December 25^{th};
W_t^{23}	January 1^{st};
W_t^{24}	December 24^{th} (if not a bridge day);
W_t^{25}	Bridge day: Monday before a holiday or Friday after a holiday;
	Other effects:
W_t^{26}	August weekend trend 1: number of days since end of July;
W_t^{27}	August weekend trend 2: number of days since 2nd half August;
W_t^{28}	Dummy variable to indicate daylight saving period.

The summer holiday period in France has a pronounced effect on electricity loads. The load levels decrease significantly in this period, since many production facilities are not operating at their full capacities and families live more outdoors. The load level decrease is gradual, and can be characterized as follows. The differences in load levels between regular weekdays and weekends decrease progressively during the first half of the summer holiday period, and increase during the second half of the summer holidays. We model this effect with the following two variables. The first variable W_t^{26} is always zero except in the weekends of the last days of July and the first two weeks of August, when it equals the number of days since the last Friday in July. The second variable W_t^{27} is always zero except in August on weekend-days after the first two weeks of the month, when it equals the number of days since the last Friday in the second week of August. The last dummy variable of the model, W_t^{28}, is created for the daylight saving

period. It distinguishes periods of winter-time and summer-time.

The values of these 28 calendar variables are collected in the vector

$$W_t = \begin{pmatrix} W_t^1 & W_t^2 & \dots & W_t^{28} \end{pmatrix}'.$$

The unknown regression coefficients for the total calendar effect are in the vector

$$\gamma_{t,i} = \begin{pmatrix} \gamma_{t,i}^1 & \gamma_{t,i}^2 & \dots & \gamma_{t,i}^{28} \end{pmatrix}'.$$

We allow all these coefficients to change over time. However, it has been clear from the beginning of this study that the variables concerned with Christmas, New Year and daylight saving (W_t^j for $j = 22, 23, 24, 28$) cannot be made time-varying, since these variables refer to yearly events and our data set spans a limited number of years. We therefore have

$$\begin{cases} \gamma_{t+1,i}^j &= \gamma_{t,i}^j + e_{t,i}^j, & j = 1, \dots, 21, 25, 26, 27, \\ \gamma_{t,i}^j &= \gamma_i^j, & j = 22, 23, 24, 28, \end{cases} \tag{2.9}$$

where the disturbance $e_{t,i}^j \sim N(0, \sigma_{e^j,i}^2)$ is serially uncorrelated and can only be correlated with $e_{t,m}^j$ for $i \neq m = 1, \dots, k$, $j = 1, \dots, 21, 25, 26, 27$ and $t = 1, \dots, n$.

2.3.4 Vector representation of model with correlated errors

The time-varying regression model (2.1) can be formulated in vector form. Define $y_t = (y_{t,1}, \dots, y_{t,k})'$ as the vector of hourly electricity loads for day t. The model for y_t is given by

$$y_t = f_t + X_t^* \beta_t + W_t^* \gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \Sigma_\varepsilon), \tag{2.10}$$

where $f_t = (f_{t,1}, \dots, f_{t,k})'$ is the vector of trends, as modelled in (2.2), for $t = 1, \dots, n$. The regression effects are represented by the parameter vectors β_t and γ_t with (fixed and time-varying) regression coefficients and the matrices X_t^* and W_t^*, which consist of k rows with the explanatory variables in $X_{t,i}$ ($i = 1, \dots, k$) and W_t, respectively. The specification of the coefficient vector β_t is implied by (2.5) and γ_t is implied by (2.9). The disturbance vector $\varepsilon_t = (\varepsilon_{t,1}, \dots, \varepsilon_{t,k})'$ is serially uncorrelated. The variance matrix Σ_ε is possibly a full matrix, such that the k equations in (2.10) can be correlated with each other. This also applies to the trends in f_t and to each time-varying parameter in β_t and γ_t. The disturbance vectors driving these multivariate dynamic processes have mean zero and full variance matrices. It is natural to assume that the regression effects for different hours of the day have similar or related impacts on the electricity loads for these hours. Hence we expect that these disturbances for different hours will be correlated.

In the case of $k = 24$, variance matrices become relatively large, and this obviously leads to many unknown parameters. Therefore, various restrictions on the variance matrices may need to be imposed when k is high. In our empirical study below, we will assume that Σ_ε is diagonal. The hourly periodic model in which we allow disturbances associated with the stochastic processes for trend components and/or time-varying regression coefficients for different hours to be correlated is a novelty in the area of electricity load forecasting. We next discuss estimation and signal extraction.

2.3.5 Estimation, signal extraction and forecasting

The multivariate model (2.10) with the dynamic processes for f_t and the elements in β_t and γ_t can be framed in a linear Gaussian state space model, given by

$$y_t = Z_t\alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \qquad t = 1, \ldots, n, \tag{2.11}$$

where the vector of hourly electricity loads y_t and the disturbance vector ε_t are the same as in (2.10). The state vector α_t contains the trend components $f_{t,i}$ and $g_{t,i}$ from (2.2), the (partly) time-varying regression coefficients $\beta_{t,i}$ from (2.5) and $\gamma_{t,i}$ from (2.9). The dynamic processes of the trend components and the time-varying regression coefficients can be generally represented by the vector Markov process (or Vector Autoregressive process) for α_t in (2.11) given by

$$\alpha_t = \begin{pmatrix} f_t' & g_t' & \beta_t' & \gamma_t' & \lambda' \end{pmatrix}',$$

where $g_t = (g_{t,1}, \ldots, g_{t,k})'$ is the vector of slope (or growth) terms associated with $f_{t,i}$ in (2.2) and λ is the vector with elements λ_i^j in (2.5) for $j = 1, 2$ and $i = 1, \ldots, k$. The (partly) time-varying system matrices Z_t, T_t and R_t are fixed and known. For our model (2.10), we have

$$Z_t = \begin{bmatrix} I & 0 & X_t^* & W_t^* & 0 \end{bmatrix}, \qquad T_t = \begin{bmatrix} I & I & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & X_t^+ \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix},$$

where the matrix X_t^+ consists of zeroes and elements in $X_{t,i}^j$ to capture the terms $\lambda_i^j X_{t,i}^j$, for $j = 1, 2$ and $i = 1, \ldots, k$, in β_t of (2.5). The matrices I are identity matrices with appropriate but possibly different dimensions. The selection matrix R_t consists of zeros and ones; it links the appropriate elements of η_t with α_{t+1} for $t = 1, \ldots, n$. The vector η_t contains the disturbances associated with the trends, $v_{t,i}$ and $w_{t,i}$ in (2.2), with $u_{t,i}$, the

innovations for the time-varying regression coefficients in (2.5) and with $e_{t,i}$ in (2.9) for $i = 1, \ldots, k$ and $t = 1, \ldots, n$. Since the corresponding disturbances of different hours are correlated, the variance matrix $\mathrm{Var}(\eta_t)$ is block-diagonal with $k \times k$ blocks of variance matrices.

The fixed regression coefficients are also placed in the state vector α_t. Fixed elements such as λ in α_t have the corresponding rows of R_t in (2.11) equal to zero. A more detailed discussion of how linear Gaussian time series models with regression effects can be formulated in state space is provided by Durbin & Koopman (2001, Chapter 3). Once the model is placed in state space form, the Kalman filter is used to predict the state vector α_t, and the associated smoothing algorithms produce estimates of α_t based on the whole sample. We have implemented filtering and smoothing algorithms that account for unknown (diffuse) initial conditions concerning α_1. The Kalman filter also computes the loglikelihood function that depends on the unknown fixed parameters in $\mathrm{Var}(\varepsilon_t)$ and $\mathrm{Var}(\eta_t)$. The loglikelihood maximization with respect to the unknown parameters in these variance matrices is based on the quasi-Newton optimization method and is typically a high-dimensional problem. We obtain good starting values via the Expectation Maximization (EM) algorithm, see Shumway & Stoffer (1982) and Koopman (1993) for further details. Furthermore, quasi-Newton methods rely on the score vector, which can be evaluated either numerically or analytically, see the discussion in Koopman & Shephard (1992). In our implementation, the maximization algorithm is based on the numerical score vector. We should note that the variance matrices $\mathrm{Var}(\varepsilon_t)$ and $\mathrm{Var}(\eta_t)$ are transformed such that they are always positive semi-definite and maximization takes place without further constraints.

Once the variance matrices $\mathrm{Var}(\varepsilon_t)$ and $\mathrm{Var}(\eta_t)$ have been estimated, the Kalman filter and smoothing algorithms are used for signal extraction, and allow us to draw time series plots of the estimated trend components, as well as the estimated time-varying regression coefficients (both with associated standard errors). The fixed regression coefficients can also be calculated in this way. Since the Kalman filter equations can deal with missing observations in a natural way, forecasting is also straightforward in our framework. By extending the data sample $y_1, \ldots, y_n$ with missing values for y_t with $t = n+1, n+2, \ldots$, and applying the Kalman filter to this extended sample, the forecasts are produced as by-products. Formally, we define the forecast of y_{n+l} by $F_{n+\ell}$. The confidence interval of the forecast $F_{n+\ell}$ can be computed using $\mathrm{Var}(F_{n+\ell})$. The econometric computations have been implemented for the object-oriented matrix programming environment of `Ox`, see Doornik (2006), with state space routines from `SsfPack` as described by Koopman et al. (1999).

2.4 Empirical results

In this section we report the results of the implementation of model (2.10), with $k = 2$, for a bivariate daily time series of electricity loads for the morning hour of 9 AM ($i = 1$) and the noon hour 12 PM ($i = 2$). The variance matrices for the bivariate trend components and each time-varying regression coefficient are full. Only the variance matrix $\mathrm{Var}(\varepsilon_t)$ of the irregular ε_t is taken as diagonal. The estimation is based on the first eight years of the data, from September 1, 1995 until August 31, 2003. The last year, from September 1, 2003 until August 31, 2004, is used for post-sample forecast evaluation. The sample size of the two daily time series is $n = 2,922$. The post-sample length is 366 days.

2.4.1 Estimation results

The estimated standard deviations of the disturbances in Eqs (2.2), (2.5), (2.9) and (2.10) are presented in Table 2.2 together with their ratios with respect to the estimated standard deviation of the irregular (the so-called q-ratios). In case the q-ratio is large, say $q > 1$, the component or regression coefficient varies greatly over time. Generally, the q-ratios are smaller for 12 PM than for 9 AM. This confirms the common belief that the load at noon is more predictable than the load in the morning hours. Large q-ratios are obtained for the time-varying cosine coefficients of the yearly cycle, all for weekdays. These values suggest that the yearly load cycles vary more for weekdays than for weekends. The time-varying holiday dummy coefficients have particular high q-ratios, which will lead to some inaccurate forecasts for holidays in our model. It is interesting that all q-ratios are sufficiently large that most coefficients do vary over time in our model.

The last column of Table 2.2 reports the disturbance correlations between 9 AM and 12 PM that are implied by the full variance matrix estimates. Whereas many standard deviations of the disturbances are clearly different for the two hours, the implied correlations are mostly estimated at values close to unity. The exceptions are the time-varying dummy effects for Saturdays and Sundays, with respective correlations of -0.70 and -0.30. This indicates substitution effects in the mornings of weekends. A relatively low electricity load at 9 AM is compensated for by a higher electricity load at 12 PM, and vice versa.

Table 2.3 reports the estimated λ coefficients and the fixed regression coefficient estimates of model (2.10). Almost all of the estimates are significant. The coefficients λ_i^j in (2.5) determine the importance of the nonlinearity and the yearly periodic effect

Table 2.2: Estimation results for model (2.10) for the in-sample period September 1, 1995 to August 31, 2003, with estimated standard deviations (st.dev.) of irregulars, of disturbances driving the stochastic processes, of trends (level and slope) in (2.2), and of time-varying regression coefficients in Eqs (2.5) and (2.9). The q-ratio is the standard deviation divided by the one of the irregular. The estimates are presented for both 9 AM and 12 PM. The estimated correlations (corr.) are reported in the last column.

		9 AM ($i=1$)		12 PM ($i=2$)		
Cmp / Expl	par	st.dev.	q-ratio	st.dev.	q-ratio	corr.
irregular	$\varepsilon_{t,i}$	185.4	1.000	283.2	1.000	–
level	$f_{t,i}$	52.7	0.284	52.4	0.185	1
slope	$g_{t,i}$	13.0	0.070	13.0	0.046	1
$X^1_{t,i}$ Heating	$\beta^1_{t,i}$	25.7	0.138	19.6	0.069	0.99
$X^2_{t,i}$ SmoHeating	$\beta^2_{t,i}$	95.3	0.513	68.2	0.241	1
$X^3_{t,i}$ SmoCooling	$\beta^3_{t,i}$	4.3	0.023	4.4	0.016	1
$W^1_t = a^{WD}_{1,t}$	$\gamma^1_{t,i}$	471.9	2.545	468.8	1.655	1
$W^2_t = b^{WD}_{1,t}$	$\gamma^2_{t,i}$	210.4	1.135	213.2	0.753	1
$W^3_t = a^{WE}_{1,t}$	$\gamma^3_{t,i}$	62.4	0.337	57.1	0.202	1
$W^4_t = b^{WE}_{1,t}$	$\gamma^4_{t,i}$	114.7	0.618	118.5	0.418	1
$W^5_t = a^{WD}_{2,t}$	$\gamma^5_{t,i}$	105.1	0.567	117.4	0.414	1
$W^6_t = b^{WD}_{2,t}$	$\gamma^6_{t,i}$	105.9	0.571	103.6	0.366	1
$W^7_t = a^{WE}_{2,t}$	$\gamma^7_{t,i}$	33.9	0.183	26.8	0.095	0.97
$W^8_t = b^{WE}_{2,t}$	$\gamma^8_{t,i}$	50.0	0.270	43.2	0.153	1
$W^9_t = a^{WD}_{3,t}$	$\gamma^9_{t,i}$	392.9	2.119	381.5	1.347	1
$W^{10}_t = b^{WD}_{3,t}$	$\gamma^{10}_{t,i}$	6.8	0.037	9.4	0.033	1
$W^{11}_t = a^{WE}_{3,t}$	$\gamma^{11}_{t,i}$	32.2	0.177	33.9	0.120	1
$W^{12}_t = b^{WE}_{3,t}$	$\gamma^{12}_{t,i}$	36.1	0.195	40.1	0.142	1
$W^{13}_t = a^{WD}_{4,t}$	$\gamma^{13}_{t,i}$	375.6	2.026	294.4	1.040	1
$W^{14}_t = b^{WD}_{4,t}$	$\gamma^{14}_{t,i}$	85.8	0.463	85.1	0.301	1
$W^{15}_t = a^{WE}_{4,t}$	$\gamma^{15}_{t,i}$	23.1	0.125	22.2	0.078	1
$W^{16}_t = b^{WE}_{4,t}$	$\gamma^{16}_{t,i}$	26.7	0.144	24.8	0.088	1
W^{17}_t Monday	$\gamma^{17}_{t,i}$	14.4	0.078	0.6	0.002	1
W^{18}_t Friday	$\gamma^{18}_{t,i}$	0.7	0.004	4.3	0.015	1
W^{19}_t Saturday	$\gamma^{19}_{t,i}$	59.2	0.319	7.1	0.025	-0.70
W^{20}_t Sunday	$\gamma^{20}_{t,i}$	133.6	0.721	107.1	0.378	-0.30
W^{21}_t Holiday	$\gamma^{21}_{t,i}$	559.5	3.018	465.8	1.645	1
W^{22}_t Bridge day	$\gamma^{22}_{t,i}$	126.0	0.679	90.7	0.320	1
W^{26}_t August Tr1	$\gamma^{26}_{t,i}$	6.0	0.032	5.0	0.018	1
W^{27}_t August Tr2	$\gamma^{27}_{t,i}$	110.7	0.597	89.6	0.317	1

in the time-varying regression coefficients of the two heating effects $X^j_{t,i}$ $(j = 1, 2)$. All the estimated coefficients are significant, providing evidence that temperature effects are subject to yearly (periodic) nonlinear behaviour. The fixed regression estimates show that cloud cover has a significant effect on the load, whereas the daylight saving effect is not significant. The latter result could have been expected for 12 PM. The special effects for Christmas and New Year are, unsuprisingly, highly significant.

Table 2.3: Estimation results for the lagged temperature coefficients of model (2.10) in the time-varying regression equations (2.5) for the heating and smoothed heating effects (leading to a nonlinear and yearly periodic dependence of temperature on electricity loads), and estimation results for the fixed regression coefficients of model (2.10) in equations (2.5) and (2.9) related to the cloud cover effect and to four calendar effects. The estimates are presented for both 9 AM and 12 PM.

Explanatory variable	hour	coefficient	estimate	stand.err.	t-value
$X^1_{t,i}$ Heating	9 AM	λ^1_1	6.84	1.02	6.69
	12 PM	λ^1_2	8.55	1.22	7.02
$X^2_{t,i}$ Smoothed-heating	9 AM	λ^2_1	16.96	4.42	3.84
	12 PM	λ^2_2	16.58	3.44	4.82
$X^4_{t,i}$ Cloud cover	9 AM	β^4_1	147	7.9	18.6
	12 PM	β^4_2	171	7.8	21.9
W^{23}_t December 25^{th}	9 AM	γ^{23}_1	-14028	326.1	43.0
	12 PM	γ^{23}_2	-10602	278.3	38.1
W^{24}_t January 1^{st}	9 AM	γ^{24}_1	-15629	321.8	48.6
	12 PM	γ^{24}_2	-10847	268.7	40.4
W^{25}_t December 24^{th}	9 AM	γ^{25}_1	-4486	336.8	13.3
	12 PM	γ^{25}_2	-4499	289.1	15.5
W^{28}_t Daylight saving	9 AM	γ^{28}_1	165	108.1	1.5
	12 PM	γ^{28}_2	108	104.6	1.0

2.4.2 In-sample signal extraction: trends and time-varying coefficients

Based on the parameter estimates in Table 2.2, we apply the Kalman filter and smoothing algorithms to produce smoothed estimates of the state vector α_t that contains the trend components and the time-varying coefficients. Our time series plots of these estimates start only in January 1, 1997, because the initialization period for the filtering and smoothing is somewhat unstable, due to dummy variables that have zeroes for a long period of time (holiday effects).

Figure 2.3 shows the estimated local linear trends for 9 AM (a) and 12 PM (b); they appear to be very smooth. They indicate that all other systematic effects in the electricity load have been captured by model (2.10). Even over a long period of several years, the underlying trends show no structural changes.

Figure 2.4 presents the time-varying regression coefficients for heating, smoothed heating and smoothed cooling at 9AM - panels (a) (c) and (e) - and 12 PM - panels (b), (d) and (f). The heating coefficients exhibit a seasonal pattern, most notably for the heating effect in displays (a) and (b). Here the coefficient is largest for the winter period, though it increases from September onwards and decreases from February onwards. Because temperature values also have a yearly cycle, this confirms the nonlinear and periodic

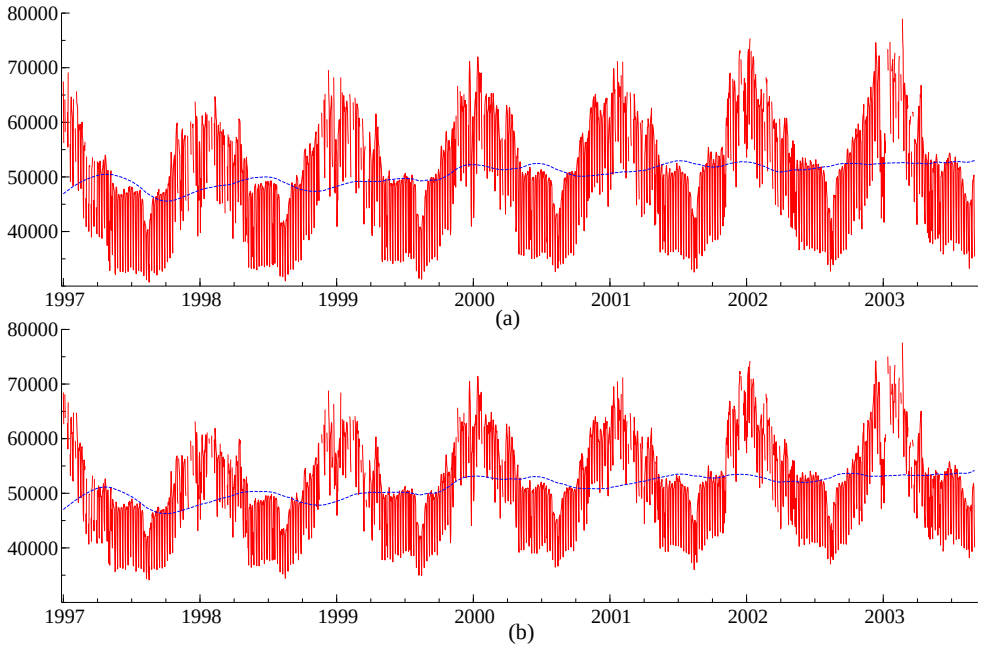

Figure 2.3: French national hourly electricity load in MWh from January 1, 1997 to August 31, 2003, and smoothed estimates of the stochastic trend in model (2.10): (a) at 9 AM, stochastic trend $f_{t,1}$; (b) at 12 PM, stochastic trend $f_{t,2}$.

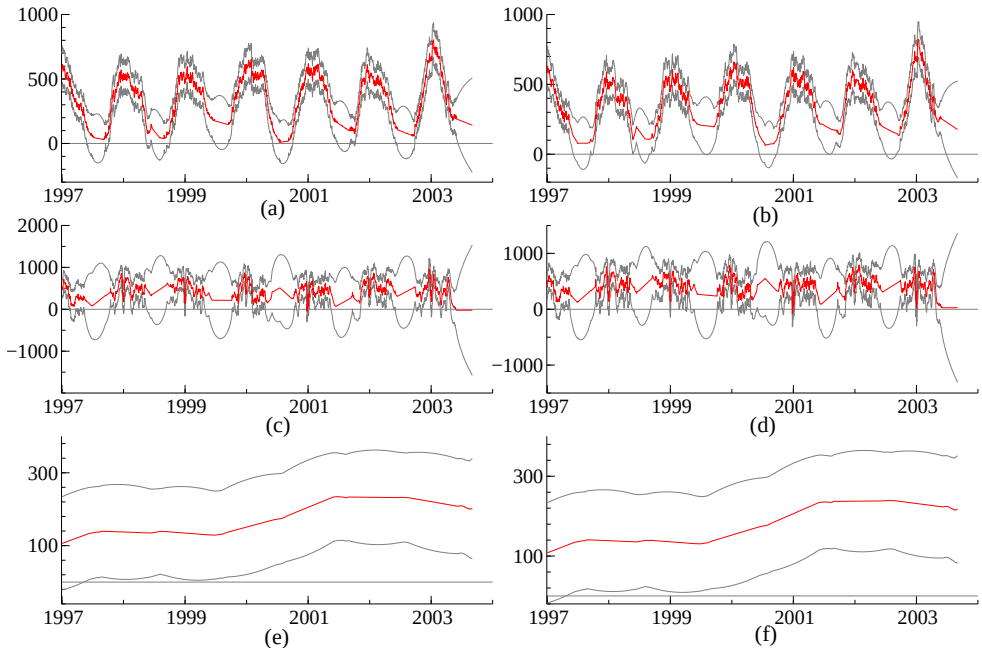

Figure 2.4: Stochastic temperature coefficients for 9 AM and 12 PM: smoothed estimates and associated 95% confidence intervals. Coefficient for heating degrees (a) at 9 AM ($\beta^1_{t,1}$) and (b) at 12 PM ($\beta^1_{t,2}$) ; Coefficient for smoothed-heating degrees (c) at 9 AM ($\beta^2_{t,1}$) and (d) at 12 PM ($\beta^2_{t,2}$) ; Coefficient for smoothed-cooling degrees (e) at 9 AM ($\beta^3_{t,1}$) and (f) at 12 PM ($\beta^3_{t,2}$) .

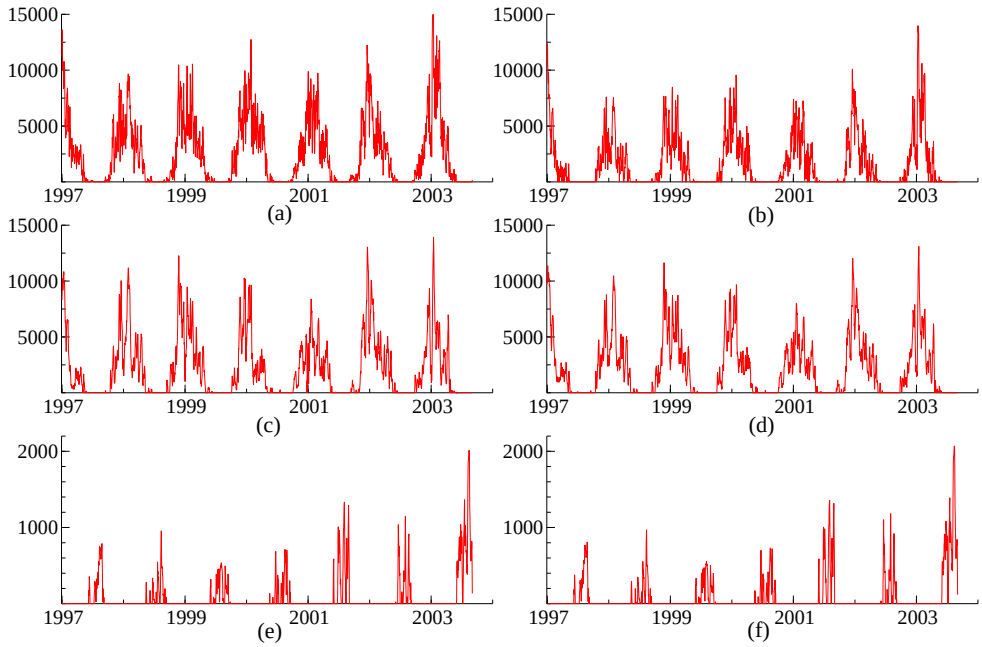

Figure 2.5: Stochastic regression effect for heating degrees (a) at 9 AM ($\beta^1_{t,1}X^1_{t,1}$) and (b) at 12 PM ($\beta^1_{t,2}X^1_{t,2}$) ; Stochastic regression effect for smoothed-heating degrees (c) at 9 AM ($\beta^2_{t,1}X^2_{t,1}$) and (d) at 12 PM ($\beta^2_{t,2}X^2_{t,2}$) ; Stochastic regression effect for smoothed-cooling degrees (e) at 9 AM ($\beta^3_{t,1}X^3_{t,1}$) and (f) at 12 PM ($\beta^3_{t,2}X^3_{t,2}$).

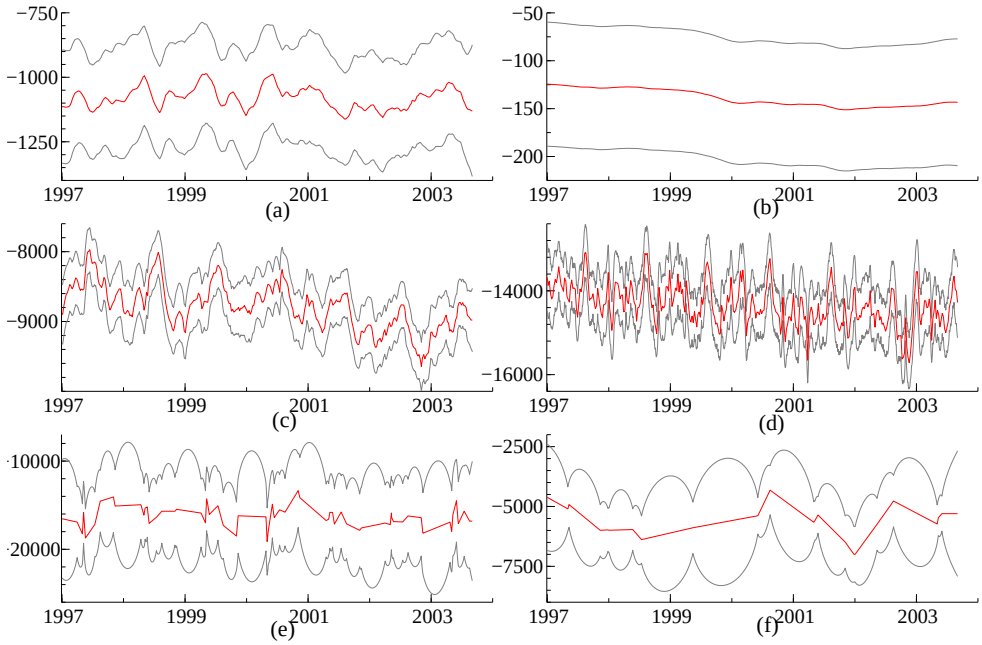

Figure 2.6: Time-varying day-type coefficients for 9 AM: smoothed estimates and associated 95% confidence intervals. (a) Coefficient for Mondays, $\gamma^{17}_{t,1}$; (b) Coefficient for Fridays, $\gamma^{18}_{t,1}$; (c) Coefficient for Saturdays, $\gamma^{19}_{t,1}$; (d) Coefficient for Sundays, $\gamma^{20}_{t,1}$; (e) Coefficient for Holidays, $\gamma^{21}_{t,1}$; (f) Coefficient for bridge days, $\gamma^{25}_{t,1}$.

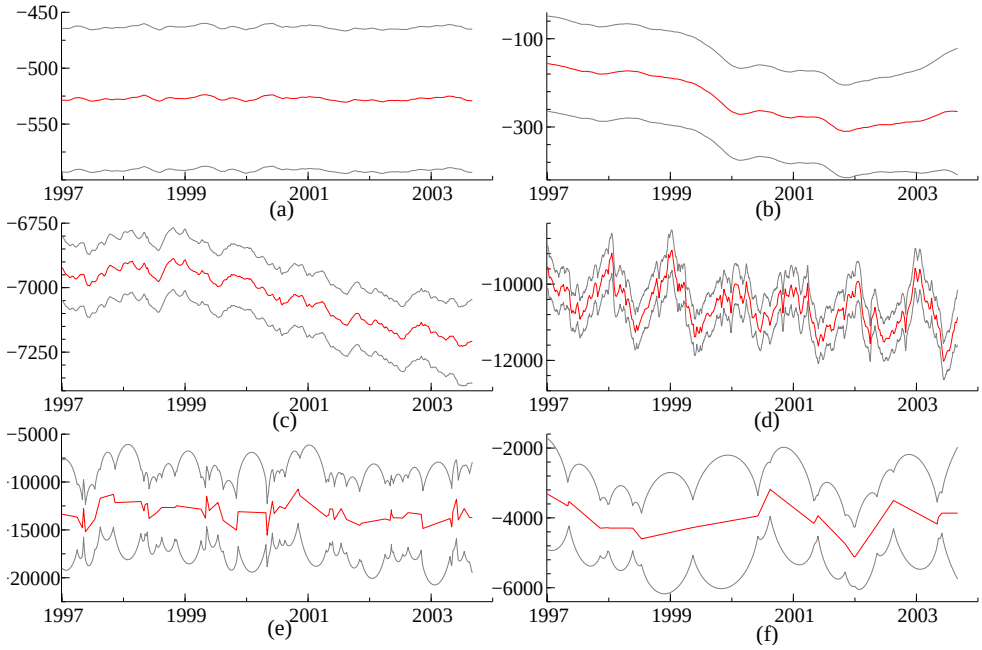

Figure 2.7: Time-varying day-type coefficients for 12 PM: smoothed estimates and associated 95% confidence intervals. (a) Coefficient for Mondays, $\gamma_{t,2}^{17}$; (b) Coefficient for Fridays, $\gamma_{t,2}^{18}$; (c) Coefficient for Saturdays, $\gamma_{t,2}^{19}$; (d) Coefficient for Sundays, $\gamma_{t,2}^{20}$; (e) Coefficient for Holidays, $\gamma_{t,2}^{21}$; (f) Coefficient for bridge days, $\gamma_{t,2}^{25}$.

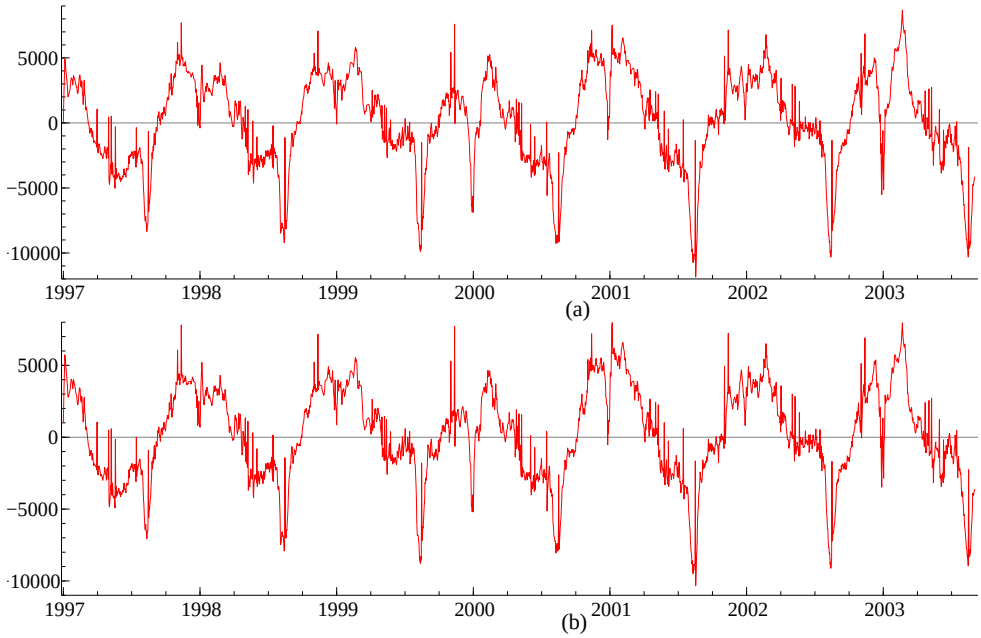

Figure 2.8: Stochastic yearly patterns in the hourly loads. Smoothed estimates of the total effect of yearly Fourier series for weekdays and weekends plus the effect of daylight saving and August trends for weekends. (a) Yearly pattern at 9 AM, $\gamma_1^{28}W_t^{28} + \sum_{j\in\{1,..,16,26,27\}} \gamma_{t,1}^{j}W_t^{j}$; (b) Yearly pattern at 12 PM, $\gamma_2^{28}W_t^{28} + \sum_{j\in\{1,..,16,26,27\}} \gamma_{t,2}^{j}W_t^{j}$;

nature of this time-varying coefficient. During the summer period, the heating coefficients are implicitly interpolated because the corresponding explanatory variables are zero, giving no information about these coefficients. The associated standard errors are higher in the summer as a result. The seasonal patterns of the heating coefficients are an interesting and novel feature of our model.

The cooling regression coefficients in panels (e) and (f) of Figure 2.4 increase slowly over time, with noticeable changes between the summers of 1999, 2000 and 2001. These increases of the cooling effect may be attributed to the growing number of installations of cooling utilities for businesses in France, especially from 1999 onwards. The cooling coefficients are implicitly interpolated during colder periods, when the explanatory cooling variable is zero. Since the number of days with a smoothed temperature larger than 18^oC is relatively small, and the significance levels of the cooling coefficients are lower than those of the heating coefficients, the standard errors are relatively constant through the summer and winter periods. It is a satisfactory empirical finding that the time-varying cooling coefficients show a clear upward trend in the period when air conditioning began to be used more intensively.

Figure 2.5 shows the estimated time-varying heating and cooling effects on the electricity load, that is, $\beta^j_{t,i}X^j_{t,i}$ for $j = 1, 2, 3$ in panels (a), (c), (e) for 9 AM, i.e. $i = 1$ (respectively (b), (d), (f) for 12 PM, i.e. $i = 2$). Naturally, the heating effects on the load are most pronounced in the winter periods. The smoothed heating effects in panels (c) and (d) have a clearer impact on the load than the actual heating effects in panels (a) and (b). The cooling effects in panels (e) and (f) have a lower impact on the load. However, the alternating heating and cooling effects in winter and summer periods are clear from Figure 2.5.

Figures 2.6 and 2.7 show the estimated effects for the different day types of (a) Mondays, (b) Fridays, (c) Saturdays, (d) Sundays, (e) holidays and (f) bridge days, at 9 AM and 12 PM, respectively. The effects are negative for all these day types and therefore the levels for those days are lower than on the regular (default) days of Tuesday, Wednesday and Thursday, when there are more business activities. For example, the Monday effect is -1100 MWh at 9 AM but only -525 MWh at 12 PM. For the other day type effects, the differences between the two hours are smaller. In general, the day type effects for 9 AM are stronger than for noon. However, the Friday effect for 12 PM is stronger than for 9 AM and it becomes stronger over time: from -125 MWh to around -250 MWh at the end of the summer 2003, see Figure 2.7, panel (b). This change may be explained by the decrease in the official number of working hours in France.

The day type effects are, not surprisingly, most pronounced for weekends and holidays.

The Saturday effect decreases slowly from around −8000 to −9500 MWh at 9 AM. A similar long-term decrease from −7000 to −7250 MWh is observed at 12 PM, but this change is much smoother over time; compare also the q-ratios for W_t^{19} in Table 2.2. The Sunday effect at 9AM varies between −13000 MWh and −15500 MWh. The strong variation in the Sunday effects may be due to the various holidays that occur around a Sunday. Special modelling of such effects may need to be considered. The significant Bridge day effect turns out to be fairly constant over time. The holiday effects are as important as the Sunday effect, and are also relatively constant.

Figure 2.8 presents the global yearly effects for (a) 9 AM and (b) 12 PM. The yearly effect consists of the impact via the time-varying Fourier coefficients (separated for weekdays and weekends), together with the daylight saving (fixed effect) and August trend effects, i.e. $\gamma_i^{28} W_t^{28} + \sum_{j\in\{1,..,16,26,27\}} \gamma_{t,i}^{j} W_t^{j}$, $i = 1, 2$. The time-varying structure of the model enables the coefficients to adapt to periods with a fast change in the yearly pattern, especially at the end of the year and in August. The result is a more parsimonious model, since it avoids the inclusion of more special dummy variables to capture these specific effects. This adaptiveness is captured by large q-ratios for some of the Fourier coefficients in Table 2.2.

2.4.3 In-sample diagnostics

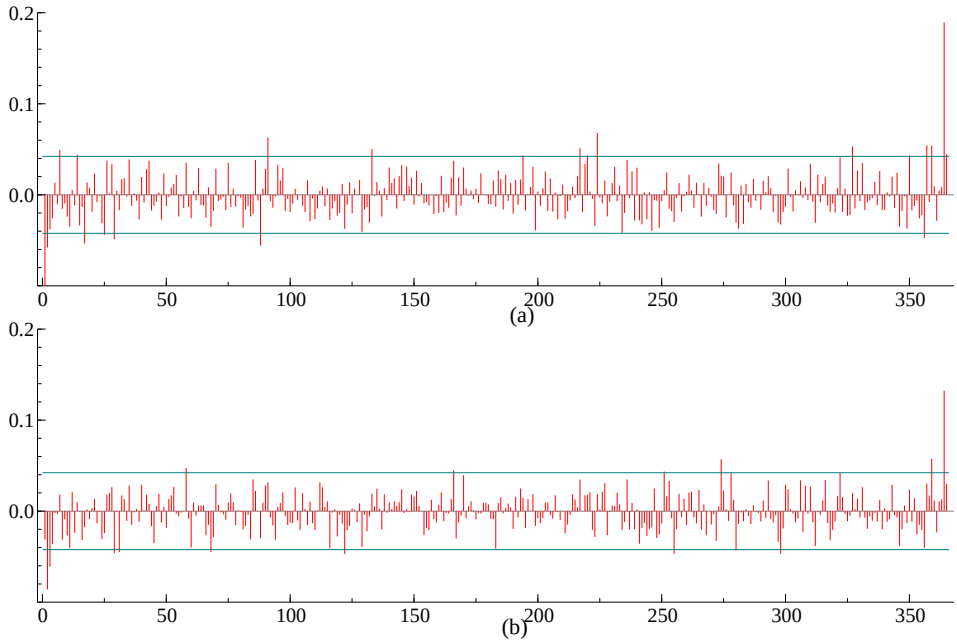

Figure 2.9: Empirical ACFs of in-sample scaled one-day ahead prediction errors (standardised) for lags $1, \ldots, 366$, at (a) 9 AM ($i = 1$) and at (b) 12 PM ($i = 2$).

The standardised prediction errors are obtained from the Kalman filter. When the model (2.10) is well-specified, the standardised prediction errors are serially independent and normally distributed. Various diagnostics can be used to check whether the prediction errors can be regarded as independent random deviates from a standard normal distribution. Since our main focus is on forecasting, we concentrate on the dynamic features of the prediction errors. A particularly interesting diagnostic is the sample autocorrelation function of the (in-sample) standardised prediction errors. Figure 2.9 displays the correlogram for lags 1 up to 365 for (a) 9 AM and (b) 12 PM. The correlations are mostly within the 95% confidence bands, which is satisfactory. For low lags, the correlations are outside the interval, but we view these values as acceptable. The one exception is the correlation at lag 365 (the one-year lag) which shows that the model does not capture all dynamics well with respect to the yearly cycle.

2.5 Forecasting performance

The forecasting performance of model (2.10) is investigated for the post-sample observations from September 1, 2003 to August 31, 2004, using the parameter estimates of Section 2.4. Since model (2.10) includes weather variables as explanatory variables, the forecasting accuracy can be based either on realised hourly values of the weather variables or on their one-day ahead forecasts. The former may be preferred to avoid having to discuss external inaccuracies due to weather forecast errors, while the latter may be preferable in that it allows us to compare models as they would be used in real situations. Since we have both the actual temperatures and one-day-ahead temperature forecasts available in our database, we evaluate the forecasts of our model for both situations. Parameter estimation and Kalman filter updating are not affected by using temperature forecasts, since at EDF the realised temperature is always available on the next day. The one-day-ahead prediction errors based on the realised temperature can then be computed and the likelihood function adjusted. We do not have forecasts for the cloud cover in our database, so here we use realised data only.

We use the mean absolute percentage forecast error (MAPE), the root mean squared forecast error (RMSE), and the mean percentage forecast error (MPE) to assess forecasting performance. For hour i, they are given by:

$$\begin{aligned} \text{MAPE}(i) &= N^{-1} \textstyle\sum_{t=1}^{N} 100 |E_{t,i}^h \,/\, y_{t,i}|, \\ \text{RMSE}(i) &= \sqrt{N^{-1} \textstyle\sum_{t=1}^{N} (E_{t,i}^h)^2}, \\ \text{MPE}(i) &= N^{-1} \textstyle\sum_{t=1}^{N} 100 E_{t,i}^h \,/\, y_{t,i}, \end{aligned} \tag{2.12}$$

where $F_{t,i}^h$ and $E_{t,i}^h = y_{t,i} - F_{t,i}^h$ are the actual forecast and the forecast error (not stan-

dardised), respectively, at day t and hour i for $t = n+1, \ldots, n+N$ and $i = 1, 2$ with N being the number of available forecasts, and h being the forecasting horizon, in our case $h = 1, \ldots, 7$.

We first analyse the results of one-day ahead forecasts for the hours of 9 AM and 12 PM. This is particularly interesting because maximum likelihood estimation optimizes the one-step-ahead prediction errors. We then look at forecasting the hourly loads up to 7 days ahead for 9 AM and 12 PM. Finally, we present and analyse one-day-ahead forecasts for all 24 hours based on one univariate model and two bivariate models.

2.5.1 One-day-ahead forecasts for 9 AM and 12 PM

Figure 2.10 presents the one-day ahead forecast errors for the electricity load at (a) 9 AM and (b) 12 PM as well as (c) their standard errors for non-EJP days. On the whole, the forecasts seem unbiased. The largest forecast errors correspond to holidays (in November, December, May, July and August). Standard errors are large during winter periods and weekends, but particularly large values are obtained for specific holidays and August weekends. These large standard errors are typically associated with effects for which only a few observations are available for estimation.

Figure 2.11 presents the correlogram of the daily prediction errors for the post-sample observations at (a) 9 AM and (b) 12 PM. The correlations for lags 1 up to 7 are somewhat larger than the in-sample correlogram, but they are not significant. However, we do not find any pattern that leads us to believe that we have missed a structural dynamic feature in the time series. The correlogram values for higher order lags vanish to zero, since the number of values that can be used to compute the higher lag correlations becomes smaller (the post-sample period consists of 366 days). Therefore we cannot use Figure 2.11 to comment on the long-term forecasting ability of the model.

Table 2.4 presents the overall MAPEs for one-day ahead forecasts of non-EJP days for 9 AM and 12 PM. The MAPEs for our full model are 1.34% and 1.31% when using realised temperatures, and 1.44% and 1.50% when using temperature forecasts. To place these measures in perspective, we also present the results for separate subperiods, and consider four benchmark models. The first benchmark model is the weekly random walk (RW). This forecast method was one of the best benchmarks in the study by Taylor & McSharry (2007). Therefore we also report the MAPE for the RW. The basic forecast function of the RW in our study is simply $F^h_{t,i} = y_{t-7,i}$, $h = 1, .., 7$. The observed load of a week ago at the same hour is the forecast for today. We take the value of two weeks ago if there was a holiday one week ago. Various problems arise with special days

including holidays. We have deleted these forecasts for the RW since we only need the RW to serve as a benchmark. We consider three restricted variants of our general model as additional benchmarks. In summary, we consider five different forecasting models:

RW: Weekly random walk;

Reg: Model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = \sigma_{u^j,i} = \sigma_{e^j,i} = 0$ in Eqs (2.2), (2.5) and (2.9) ;

Univ: A univariate version of model (2.10), i.e. $k = 1$ (trend components and regression coefficients time-varying);

TVR: Model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = 0$ in (2.2), (regression coefficients time-varying);

TTR: Model (2.10) as it is, with $k = 2$ (trend components and regression coefficients time-varying).

The overall forecasting performances for non-EJP days of all five models are reported in Table 2.4, with separate evaluations by day type.

Table 2.4: One-day ahead forecasting results by day type and by hour for the post-sample period September 1, 2003 to August 31, 2004, using the actual (left) and forecast (right) temperature. The MAPE in (2.12) is reported for five different models: a weekly random walk (RW); a regression model (Reg) , i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = \sigma_{u^j,i} = \sigma_{e^j,i} = 0$ in (2.2), (2.5) and (2.9); a univariate model (Univ), i.e. model (2.10) with $k = 1$; a time-varying regression model (TVR), i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = 0$ in (2.2); and a local linear trend plus time-varying regression model (TTR) given in (2.10) with $k = 2$. The total MAPE for all 343 non-EJP days is also reported as well as the MPE in (2.12) for EJP days. The RW model does not necessarily produce forecasts for all N observations.

				MAPE (real temperature)				MAPE (forecast temperature)			
Hour	Day type	N	RW	Reg	Univ	TVR	TTR	Reg	Univ	TVR	TTR
9	Default	132	4.63	2.23	0.93	**0.92**	0.92	2.26	**1.02**	1.06	1.05
	Monday	47	4.88	2.66	1.39	1.39	**1.36**	2.73	1.45	1.47	**1.44**
	Friday	48	4.50	1.93	**1.01**	1.02	1.01	1.86	**1.06**	1.09	1.10
	Saturday	49	5.41	2.38	1.42	**1.22**	1.23	2.38	1.35	**1.22**	1.27
	Sunday	51	6.39	2.73	1.69	1.59	**1.55**	2.77	1.87	1.71	**1.68**
	Special	16	–	9.03	5.49	**5.33**	5.40	9.16	5.47	5.48	**5.44**
	Total	343	5.04	2.66	1.40	1.34	**1.34**	2.68	1.47	1.45	**1.44**
12	Default	132	4.51	1.90	**0.94**	0.96	0.97	1.91	**1.13**	1.18	1.19
	Monday	47	4.27	2.15	**1.35**	1.40	1.42	2.16	**1.53**	1.53	1.53
	Friday	48	4.38	1.92	1.17	1.16	**1.12**	1.89	1.43	1.33	**1.30**
	Saturday	49	4.80	2.08	1.09	1.01	**1.01**	2.12	1.18	1.14	**1.13**
	Sunday	51	5.50	2.81	**1.26**	1.44	1.45	2.81	**1.54**	1.69	1.70
	Special	16	–	6.61	**4.23**	4.68	4.82	6.95	**4.33**	4.97	5.03
	Total	343	4.66	2.34	**1.25**	1.30	1.31	2.34	**1.44**	1.50	1.50
				MPE(real temperature)				MPE (forecast temperature)			
Hour	Day type	N	RW	Reg	Univ	TVR	TTR	Reg	Univ	TVR	TTR
9	EJP	23	–	-1.52	-4.25	-4.87	-4.77	-1.71	-4.52	-5.21	-5.13
12	EJP	23	–	-2.36	-4.55	-4.64	-4.55	-3.66	-6.36	-6.15	-6.09

Table 2.5: One-day-ahead forecasting results by month and by hour for the post-sample period September 1, 2003 to August 31, 2004, using the actual (left) and forecast (right) temperature. The MAPE in (2.12) is reported for five different models: a weekly random walk (RW); a regression model (Reg) , i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = \sigma_{u^j,i} = \sigma_{e^j,i} = 0$ in (2.2), (2.5) and (2.9); a univariate model (Univ), i.e. model (2.10) with $k = 1$; a time-varying regression model (TVR), i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = 0$ in (2.2); and a local linear trend plus time-varying regression model (TTR) given in (2.10) with $k = 2$. The RMSE in (2.12) is reported for temperature forecast errors (^oC). The RW model does not necessarily produce forecasts for all N observations. Overall results for all 343 non-EJP days are in Table 2.4.

			MAPE (real temperature)					RMSE	MAPE (forecast temperature)			
Hour	Month	N	RW	Reg	Univ	TVR	TTR	^oC	Reg	Univ	TVR	TTR
9	January	25	5.84	3.03	1.97	1.67	**1.65**	0.58	3.07	1.93	1.58	**1.53**
	February	20	5.76	**0.98**	1.30	1.50	1.49	0.77	**0.92**	1.31	1.46	1.46
	March	27	10.79	2.35	1.16	1.17	**1.12**	1.16	2.21	**1.47**	1.71	1.69
	April	30	7.59	1.15	1.24	**1.13**	1.27	1.00	**1.37**	1.52	1.40	1.47
	May	31	5.38	3.24	2.33	**1.86**	2.02	0.75	3.35	2.20	**1.84**	2.03
	June	30	1.14	1.79	**0.60**	0.64	0.63	0.37	1.81	**0.61**	0.64	0.62
	July	31	1.16	2.22	**1.02**	1.12	1.06	0.61	2.24	**1.03**	1.14	1.08
	August	31	5.36	7.07	2.22	1.99	**1.97**	0.56	7.07	2.21	1.99	**1.97**
	September	30	1.78	2.26	**0.66**	0.72	0.72	1.05	2.31	**0.69**	0.74	0.75
	October	31	6.40	2.00	1.26	1.02	**1.00**	1.06	1.92	1.24	1.02	**1.01**
	November	30	5.24	1.65	1.30	1.45	**1.27**	1.00	1.68	**1.48**	1.72	1.52
	December	27	5.70	3.64	**1.78**	1.98	2.04	0.76	3.67	**1.97**	2.23	2.29
12	January	25	5.57	2.75	1.83	1.77	**1.73**	1.19	2.36	1.85	**1.68**	1.74
	February	20	6.82	**1.06**	1.16	1.26	1.31	1.82	**2.02**	2.23	2.08	2.10
	March	27	10.05	1.67	**1.02**	1.17	1.16	1.55	**1.51**	1.60	1.76	1.71
	April	30	6.80	**1.12**	1.41	1.35	1.48	0.67	1.38	**1.38**	1.52	1.55
	May	31	4.50	2.75	**1.72**	1.76	1.94	0.70	2.70	**1.75**	1.82	1.98
	June	30	1.26	2.04	0.66	**0.61**	0.64	0.61	2.04	0.66	**0.61**	0.64
	July	31	1.15	1.89	**1.05**	1.19	1.12	0.80	1.89	**1.05**	1.19	1.12
	August	31	4.71	5.85	**1.53**	1.62	1.60	0.82	5.85	**1.53**	1.62	1.60
	September	30	0.97	2.24	**0.69**	0.75	0.72	0.94	2.24	**0.69**	0.75	0.73
	October	31	5.97	2.06	**1.03**	1.09	1.08	0.97	2.03	**1.13**	1.17	1.16
	November	30	5.18	1.20	1.27	1.34	**1.19**	1.45	**1.16**	1.87	2.01	1.90
	December	27	4.80	2.69	**1.71**	1.77	1.82	1.33	2.65	**1.96**	2.05	2.12

We first discuss the forecasting results in Table 2.5, where we report the MAPE, as defined in (2.12), for the five models separately for each month of the forecasting period, and where we compare forecasts based on the realised temperature (left part of the table) with those based on one-day-ahead temperature forecasts (right part of the table). The RW is obviously not affected by this choice of temperature values. The number of forecast errors for each month N is indicated in the third column of Table 2.5. The number of forecasts produced by RW may be less than N. The ninth column gives the RMSEs of the one-day ahead temperature forecasts. The most interesting aspects of Table 2.5 are:

- The RW performs rather poorly for all months and both hours, except for June, July and (in a less pronounced way) September. However, the RW only outperforms the fixed regression (Reg) model for these months.
- The univariate model, Univ, outperforms both RW and Reg, with the exception of February, where Reg is the best for both 9 AM and 12 PM.
- The TVR and TTR models outperform the three other models overall. Time-varying trends do not necessarily lead to better forecasts. Most of the differences between the MAPEs of TTR and TVR models are small.
- The forecasting results for TTR and TVR are disappointing for January, May, August and December. The most accurate forecasts are for June and September. Forecasting results for the two hours of the day studied are comparable.

Comparing the forecasting accuracy using realised and forecasted temperatures, we find the following. With respect to the effect of temperature forecast errors on load forecast errors, we note that the one-day-ahead temperature forecast RMSE of 1.16^oC in March at 9 AM seems to generate a large MAPE for TVR and TTR, and, to a lesser extent, for Univ. A similar effect is observed in November and December. In April at 9 AM and in February at 12 PM, Univ is more affected by the temperature forecast errors than TVR and TTR. These findings illustrate the importance of temperature forecast accuracy. However, temperature forecast errors have a smaller impact on the forecasting accuracy in the summer months. This confirms that cooling effects have a smaller impact on the electricity load than heating effects. Forecasts from time-varying regression models still outperform those of Reg and RW when the forecasts are based on one-day ahead temperature forecasts.

Table 2.4 presents the MAPEs for each day type for 9 AM and 12 PM. It reports the forecast results for all special days (holidays, bridge days) together in one category. For obvious reasons, forecasts for the RW model are missing for these days. Interesting findings from Table 2.4 are:

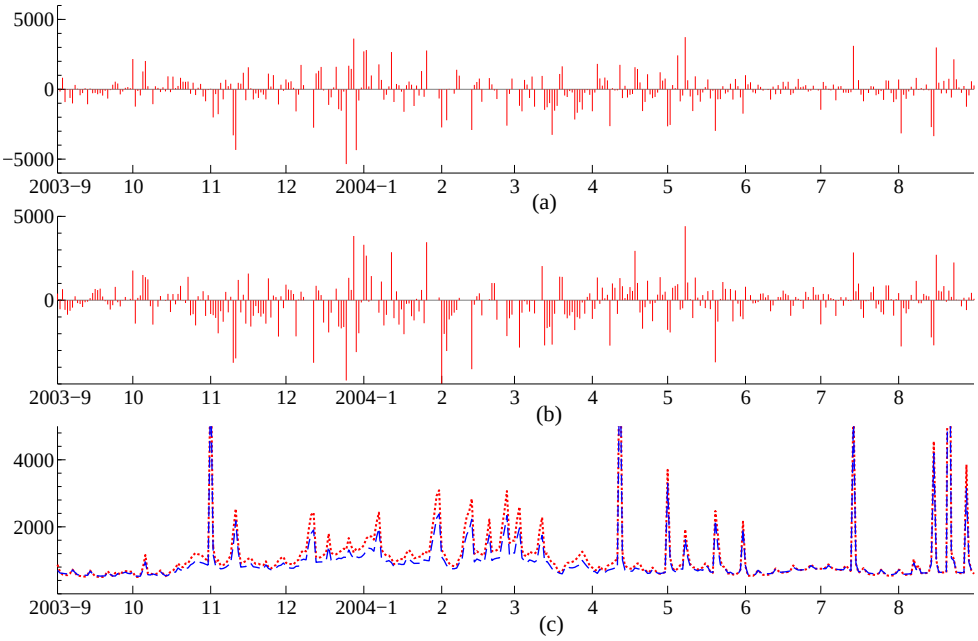

Figure 2.10: Post-sample hourly one-day-ahead forecast errors $F^1_{t,i}$ based on weather forecasts at (a) 9 AM ($i = 1$) and at (b) 12 PM ($i = 2$), with (c) associated standard errors, for $t = n + 1, .., n + 366$, .

- The holiday loads are the most difficult to forecast, since there are fewer of them, and the observed loads vary more than on other days.
- Overall the forecast accuracy measures for Univ, TVR and TTR are smaller than those for Reg and RW. More specifically, the forecasts of TVR and TTR outperform those of Univ at 9 AM, while at 12 PM, the forecasts of the univariate model Univ generally outperform those of TVR and TTR.
- Default days and Fridays (Saturdays) are forecasted most accurately at 9 AM (12 PM). Loads on Mondays and Sundays are more difficult to predict.
- The forecast accuracies obtained for 9 AM and 12 PM are comparable.

Qualitatively, these findings do not alter when forecasts are based on realised temperatures rather than one-day-ahead temperature forecasts. For completeness, Table 2.4 also reports the mean percentage forecast errors (MPE) for EJP days. The EJP days are treated as missing for the estimation of parameters in Section 2.4, but the Kalman filter can still produce forecasts for these days. The bias in forecasting the EJP days is clear, and we conclude that the model systematically over-estimates the realised electricity loads for these days.

Overall we are satisfied with the post-sample forecasting performance of our model. We have shown that time-varying and periodic regression effects are important in accu-

rately forecasting hourly loads.

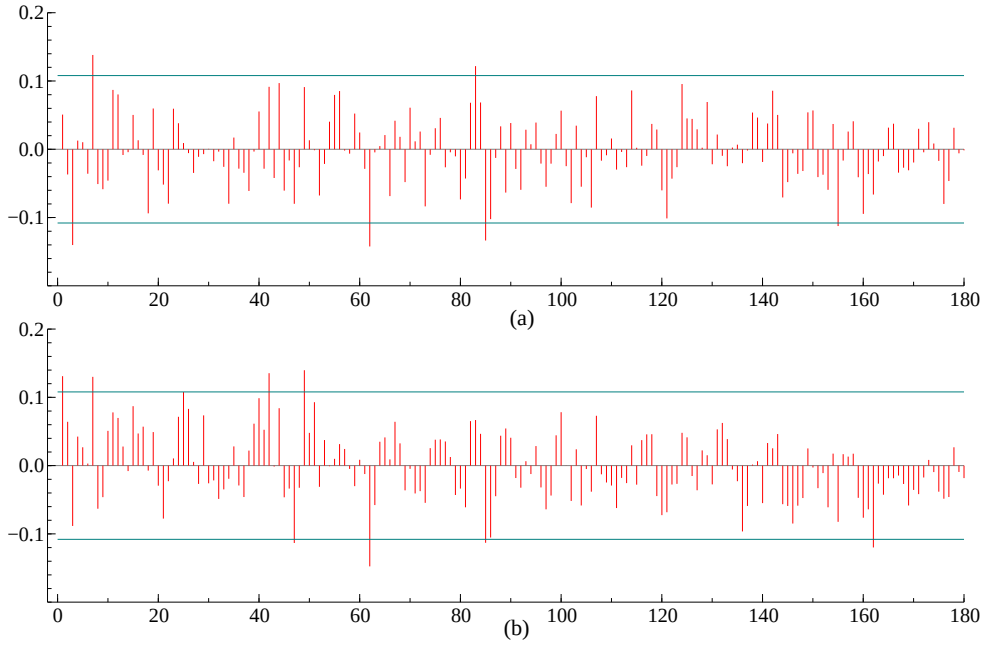

Figure 2.11: Empirical autocorrelation function for post-sample one-day-ahead forecast errors $y_{t,i} - F^1_{t,i}$ based on weather forecasts, standardised, for $t = n+1, .., n+366$, at (a) 9 AM ($i = 1$) and at (b) 12 PM ($i = 2$).

2.5.2 Multi-day ahead forecasts

Table 2.6 shows the forecast precisions for multiple day ahead forecasts (one to seven days), and for 9 AM and 12 PM. The RW has the same MAPE value for all horizons, because the one-step-ahead forecast is based on the load of a week ago. The forecasts in this table are computed using realised values of weather variables, as forecasts for the 7-day horizon were not available in our data set.

For all forecast horizons, the RW does badly in terms of MAPE, both at 9 AM and 12 PM. All time-varying regression models perform better than model Reg up to five days ahead. TTR is best at 9 AM for one to five days ahead forecasting. At 12 PM, the Univ model is best for forecasting one to six days ahead. Model Reg is best for forecasting six and seven days ahead at 9 AM and for seven days ahead at 12 PM. These results confirm that our time-varying model is primarily designed for short-term forecasting. The results for the TVR and TTR models are very similar, and the model specifications differ only in the trend component. The estimated trends for the two models have not changed much in the evaluation period; see, for example, Figure 2.3(a).

Table 2.6: Forecasting results for different forecast horizons, for the post-sample period September 1, 2003 to August 31, 2004 (total non-EJP days is 343, N decreases with the forecasting horizon). The MAPE in (2.12) is reported for five different models: a weekly random walk (RW); a regression model (Reg) , i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = \sigma_{u^j,i} = \sigma_{e^j,i} = 0$ in (2.2), (2.5) and (2.9); a univariate model (Univ), i.e. model (2.10) with $k = 1$; a time-varying regression model (TVR), i.e. model (2.10) with $k = 2$ and $\sigma_{v,i} = \sigma_{w,i} = 0$ in (2.2); and a local linear trend plus time-varying regression model (TTR) given in (2.10) with $k = 2$. The RW model does not necessarily produce forecasts for all N.

			MAPE (Real temperature)				
Hour	Horizon	N	RW	Reg	Univ	TVR	TTR
9	1 day	343	5.04	2.66	1.40	1.34	**1.34**
	2 days	342	5.04	2.68	1.92	**1.84**	1.85
	3 days	341	5.04	2.70	2.26	**2.19**	2.21
	4 days	340	5.04	2.72	2.47	**2.36**	2.39
	5 days	339	5.04	2.74	2.71	**2.57**	2.57
	6 days	338	5.04	**2.76**	2.87	2.83	2.85
	7 days	337	5.04	**2.77**	3.07	3.06	3.09
12	1 day	343	4.66	2.34	**1.25**	1.30	1.31
	2 days	342	4.66	2.32	**1.58**	1.70	1.72
	3 days	341	4.66	2.36	**1.87**	2.00	2.02
	4 days	340	4.66	2.38	**2.01**	2.11	2.14
	5 days	339	4.66	2.40	**2.10**	2.24	2.25
	6 days	338	4.66	2.41	**2.30**	2.45	2.49
	7 days	337	4.66	**2.42**	2.49	2.67	2.66

2.5.3 One-day ahead forecasts comparison for all hours

For a more general assessment of our methodology, we consider forecasting all twenty-four hours of the day based on the models

Univ : model (2.10) with $k = 1$;

TTR-1 : model (2.10) with $k = 2$, for consecutive hours $(i - 1, i)$;

TTR+1 : model (2.10) with $k = 2$, for consecutive hours $(i, i + 1)$,

where $i = 0, \ldots, 23$ and $i - 1 = -1$ refers to the last hour of the previous day.

In Table 2.7 we compare the one-day-ahead forecasting accuracy for each hour i, with $i = 0, 1, \ldots, 23$, in terms of RMSE and MAPE. The load forecasts are based on one-day-ahead temperature forecasts. Only regular days (non-holidays, non-bridge days, non-EJP days) are considered. In terms of RMSE, Univ outperforms the bivariate models for seven hours: 13 to 17, and 21 and 22. However, in terms of MAPE, Univ outperforms those models only for two hours: 1 AM and 2 PM. We prefer the MAPE over the RMSE as it is less sensitive to outliers and is easier to compare across different applications.

In view of these forecasting results, we comment on the estimated correlations in the

Table 2.7: One-day ahead forecasting results by day type and by hour for the post-sample period September 1, 2003 to August 31, 2004, using forecast temperature. Total non-special and non-EJP days is 327 (326 at 3AM). The RMSE and the MAPE in (2.12) are reported for three different models: a univariate Model (Univ), i.e. model (2.10) with $k = 1$; TTR model given in (2.10) with $k = 2$, for consecutive hours (i-1,i) (TTR-1), TTR model given in (2.10) with $k = 2$, for consecutive hours (i,i+1) (TTR+1). Total RMSE and MAPE for the 24 hours is also reported (7847 observations).

		RMSE			MAPE		
Hour	N	Univ	TTR-1	TTR+1	Univ	TTR-1	TTR+1
0	327	966	**755**	993	1.43	**1.11**	1.42
1	327	986	984	**971**	**1.40**	1.41	1.40
2	327	992	961	**957**	1.54	1.50	**1.47**
3	326	946	929	**910**	1.52	1.52	**1.50**
4	327	922	894	**846**	1.50	1.49	**1.44**
5	327	859	858	**823**	1.44	1.44	**1.38**
6	327	1028	1067	**1027**	1.56	**1.54**	1.55
7	327	1178	1170	**1164**	1.64	1.61	**1.55**
8	327	983	1016	**979**	1.32	1.32	**1.29**
9	327	958	**954**	964	1.27	1.27	**1.25**
10	327	1097	1043	**1041**	1.44	1.33	**1.31**
11	327	1128	1091	**1070**	1.43	1.37	**1.34**
12	327	1041	1041	**1032**	1.30	1.30	**1.25**
13	327	**1050**	1080	1074	1.38	**1.38**	1.39
14	327	**1025**	1088	1062	**1.39**	1.45	1.43
15	327	**1046**	1108	1057	1.47	1.53	**1.47**
16	327	**1089**	1099	1132	1.61	**1.59**	1.63
17	327	**1139**	1158	1164	1.61	1.60	**1.59**
18	327	1137	1217	**1111**	1.55	1.64	**1.54**
19	327	1057	1065	**1050**	1.46	1.48	**1.42**
20	327	942	**922**	1003	1.34	**1.32**	1.45
21	327	**795**	867	814	1.20	1.26	**1.20**
22	327	**759**	785	759	1.07	1.12	**1.06**
23	327	722	728	**680**	1.06	1.07	**0.92**
Total	7847	993	995	**987**	1.41	1.40	**1.39**

bivariate models, which we do not present here to save space. In the morning hours, the forecasts produced by the bivariate models are better than those produced by the univariate models. The estimated disturbance correlations for each morning hour are all very close to one, which may indicate that these high correlations lead to more accurate forecasts. Correlations for the disturbances for heating, smoothed heating, smoothed cooling and level components are all close to unity in all bivariate models for all 24 hours. These estimation results will be useful in the specification search for a more parsimonious forecasting model for all 24 hours.

From Table 2.7 we conclude that a bivariate modelling approach where we allow strong correlations between time-varying regression effects for different hours can improve the general forecasting performance.

2.6 Conclusion

We present a linear multivariate periodic state space model for the forecasting of hourly electricity loads. The model includes a stochastic trend component, together with fixed and time-varying regression effects. Each equation in the model is associated with a specific hour and has different coefficients and different time-varying processes which are possibly correlated through the disturbances that drive them. Kalman filter methods are used for estimation, signal extraction and forecasting.

Our linear Gaussian time series model has a relatively simple structure: it can be multivariate, and has trend components and regression effects (fixed and time-varying). The EDF data set of French national loads consists of a long time series with hourly observations for loads, temperature, cloud cover and one-day ahead temperature forecasts. We capture interesting trends and time-varying regression coefficients from our empirical study. Some of our empirical findings have been known by experts at EDF but have not been properly measured earlier. For example, the slow increase of the cooling effect on loads, the yearly patterns in the heating regression coefficients and the strong correlations between the effects for different hours (9 AM and 12 PM).

However, the main purpose of our model is short-term load forecasting. The forecasting results are satisfactory for 1, 2 and 3 days ahead. Some improvements can be made for longer forecast horizons, but it shows that the innovations from our model mostly affect the short-term dynamics. We may need to focus further on finding more appropriate dynamic specifications for the intra-yearly variations in loads. The next challenge for our periodic model-based time-varying parameter approach to forecasting loads is to extend the model to more than two hours, estimating all components simultaneously. However, larger models usually require the estimation of more parameters. This introduces more uncertainty into a model-based analysis, and may lead to less accurate forecasts. We should therefore aim to find parsimonious formulations of the multivariate model. We believe that this is feasible, because we have found that the load components of different hours are highly correlated.

Chapter 3

Dynamic factors in periodic time-varying regression models

Abstract We consider dynamic multivariate periodic regression modelling for high frequency data. The dependent univariate time series is transformed to a lower frequency multivariate time series for periodic regression modelling. For hourly series we specify 24 equations, one for each hour of the day. The regression coefficients differ across equations and vary stochastically over time. Since the unrestricted model may contain many unknown parameters, we develop a methodology within the state-space framework to model dynamic factors in the coefficients that drive the dynamics across equations. This leads to more precise estimates of the unobserved components. We present a simulation study on a basic model and compare results with a univariate benchmark model. Then, we apply our method to French national hourly electricity loads with weather variables and calendar variables as regressors. We use block diagonal specification for factor loading matrices. The model is estimated on a long dataset for groups of hours independently and we also estimate univariate models as benchmarks. We analyse the method both from the signal extraction and from the forecasting accuracy standpoint. We find that although the methodology is more restrictive than the independent modelling of each hour, the interpretation of the different dynamics that compose electricity load remains similar and that forecasing accuracy is not deteriorated, so that the dynamic factor methodology in time-varying regression coefficients is effective for hourly electricity demand. The method is a first step for the modelling of the intradaily pattern of the different components of electricity load.

3.1 Introduction

This chapter develops a method to analyse common dynamic features in multivariate time varying regression models. We adopt the method to investigate parameter changes in high frequency periodic time series. The main idea is to introduce dynamic factor models for subseries with similar characteristics. In our case, we analyse daily subseries for specific hours of each day. The aim is to find common dynamic features in time-varying regression coefficients for different periods. We discuss the implementation of a dynamic factor state space model, including time-varying coefficients for mean, seasonal and regression effects. We show that the method works for a long series of hourly electricity loads where we take account of stochastic trends, yearly cycles, calendar effects and the changing influence of temperature.

The challenge of modelling high frequency periodic time series is the detection of the recurring but persistently changing patterns within the days, weeks and years. Some patterns are more variable than others and imply different forecasting functions. Fixed patterns can be used for long forecast horizons, whereas variable patterns are more relevant for short term forecasts.

Time-varying regression models provide a convenient statistical framework to tackle this problem. Regression models in the context of time series may contain a constant, seasonal effects and other explanatory variables. When the associated regression coefficients are allowed to change over time, the result is a flexible methodology that transforms the mean into a trend function and enables tracking time-varying patterns in the seasonal effects. Furthermore, changes in the interactions between the dependent variables and their explanatory variables can be detected.

In the case of high-frequency seasonal time series, the challenge of signal extraction and forecasting is even higher. The confounding of different seasonal effects in the same time series become apparent since, for example, daily time series are subject to quarterly effects (summer, winter), day-of-the-week effects (weekday, weekend), calendar effects (Christmas, Easter) and, possibly, weather effects. Such issues become even more important when the time series to forecast corresponds to hourly measurements. An example in the context of electricity loads is given by Harvey & Koopman (1993) where hourly loads are forecasted using time-varying regression smoothing splines. Koopman & Ooms (2006) argue that forecasting high-frequency time series based on periodic time series models can be successful. In this approach, multiple seasonal and periodic effects can be disentangled in a straightforward way.

In Dordonnat et al. (2008), see also chapter 2, and in this chapter, a periodic approach

for the forecasting of hourly electricity loads is adopted. First, hourly loads are collected in a vector of daily time series. A multivariate time series model is then developed where each regression equation can be specified separately. The regression coefficients are allowed to change over time. Since a high number of coefficients in the model can be expected in practical applications, a methodology is proposed to reduce the parameter dimensions. In the case of hourly time series, it is expected that the dynamic relation between electricity demand and, say, temperature is similar at the hours of, say, 3, 4 and 5 in the morning. To impose a common factor for the corresponding time-varying parameters may even provide a more robust forecast function since information is shared amongst a set of hours.

Given the typical noisy structure of the time series, the large number of recurring effects that need to be considered in the analysis and the vast amount of available data, forecasting electricity loads is widely seen as a challenging task. A review of load forecasting methods is given by Bunn & Farmer (1985) and the references therein. A more up-to-date review of the literature on electricity load forecasting is provided by Lotufo & Minussi (1999). Taylor & McSharry (2007) discuss short-term load forecasting (one-hour to one-day ahead) using standard forecasting methods including seasonal autoregressions and exponential smoothing which are also carried out in a periodic fashion by treating each hourly load as a daily time series. Advanced methods for load forecasting are developed by Cottet & Smith (2003) who adopt Bayesian procedures for forecasting high-dimensional vectors of time series. The covariance structures in such multivariate time series are of key importance for an effective forecasting strategy. Both Smith & Kohn (2002) and Cottet & Smith (2003) take account of the correlation between hourly loads when computing their forecasts. Espinoza, Joye, Belmans & De Moor (2005) analyse many time series at different grid point separately using periodic models in order to find commonalities in the identified characteristics of the time series. In this way, common profiles in time series are formulated which form the basis for the joint forecasting of the time series.

The internal statistical model of Electricité de France is described in Bruhns et al. (2005) and is primarily developed for the French hourly load time series forecasting. An alternative model for the signal extraction and forecasting of hourly electricity loads in France is proposed in this chapter. For this purpose, a periodic model-based approach is adopted within the multivariate Gaussian state space framework. The Kalman filter and related smoothing methods are of key importance for signal extraction and forecasting. State space models are commonly used for the analysis of a wide range of statistical time series models including autoregressive integrated moving average (ARIMA) models,

regression models with fixed or time-varying coefficients and the unobserved components time series models of Harvey (1989). They can also be used to model stationary or non-stationary data. The state space framework is already adopted for forecasting French hourly electricity loads in Dordonnat et al. (2008), see also chapter 2.

Dynamic regression factors are imposed to control the dimension of time-varying coefficients and the number of unknown parameters in associated covariance matrices. This results in an effective model for signal extraction and forecasting of French hourly electricity loads. In our model we use the early econometric approach of Engle & Watson (1981), who specify a one-factor model with constant factor loadings for a multivariate time series of wage rates. We also compare our dynamic factor models with independent univariate models. We use a combination of two estimation methods. Whereas Watson & Engle (1983) compared a scoring algorithm and an EM (Expectation Maximization) algorithm, we combine the EM method of Shumway & Stoffer (1982) with Quasi-Newton methods. As the dimension of our dependent variable is relatively large, our model can be related to panel data models. However, the model selection criteria for the number of factors in panel time series which are based on an increasing cross-section and time series dimensions, as in Bai & Ng (2002), are not directly relevant. We impose the number of dynamic factors a priori.

Our approach is based on state space modelling of multivariate unobserved components models introduced in chapter 8 of Harvey (1989) and is also related to work of Peña & Poncela (2004), who analyse a state space model with nonstationary dynamic factors. They compare the theoretical precision of one-factor model predictions and corresponding univariate model forecasts. We compare our factor model with univariate models, both for signal extraction and forecasting. Alonso, Garcia-Martos, Rodriguez & Sanchez (2008) extend the model of Peña & Poncela (2004) with seasonally nonstationary factors to form a multivariate periodic model for hourly electricity prices.

Dynamic factor models are still widely used for multivariate macroeconomic data. Giannone, Reichlin & Sala (2006) compare VAR and factor models for the estimation of business cycles. Del Negro & Otrok (2008) model business cycles in a specification with time-varying factor loadings. We find little evidence of macroeconomic effects on daily electricity consumption as our hourly data set is not long enough to cover a full business cycle.

The state space methodology for nonstationary dynamic factor models has also found many applications outside economics. For example, Ortega & Poncela (2005) propose a dynamic factor model for fertility rates and Muñoz Carpena, Ritter & Li (2005) use a dynamic factor model for groundwater quality trends. However, as far as we are aware,

nearly all dynamic factors in the literature are associated with commons trends, common cycles or common seasonal factors. In this chapter we focus on the new application of the dynamic factor method in a model with common factors in stochastic regression coefficients.

The remainder of the chapter is organized as follows. Section 3.2 formulates our model and discusses the state space framework and implementation details. Section 3.3 presents a detailed simulation study to illustrate the advantages of dynamic factor modelling for time-varying-parameter estimation and therefore for signal extraction of the temperature effect on electricity loads. Our application to French national hourly electricity loads is detailed in section 3.4. Section 3.5 concludes.

3.2 Dynamic factor regression models for periodic time series

We develop a model for univariate time series subject to seasonal fluctuations associated with different seasonal periods and subject to additional periodic time-varying regression effects. We first concentrate on the shortest seasonal period S. We are specifically interested in the case of high frequency data, where S is relatively large: e.g. in the case of hourly data $S = 24$, in the case of half-hourly data, $S = 48$. Following the periodic time series literature as in Tiao & Grupe (1980), we first transform the univariate time series to an $S \times 1$ vector time series $y_t = (y_{1,t} \dots y_{S,t})', t = 1, \dots, T$. In the case of hourly data, the time series regression model for the daily vector y_t implies a periodic model for the hourly series.

The general time-varying regression model for y_t we consider is written as:

$$y_t = \mu_t + \sum_{k=1}^{K} B_t^k x_t^k + \varepsilon_t, \quad \varepsilon_t \sim IIN(0, \Sigma_\varepsilon), \quad t = 1, \dots, T, \tag{3.1}$$

where $\mu_t = (\mu_{1,t} \dots \mu_{S,t})'$ is the $S \times 1$ vector of trend components, which captures the smooth long-term evolution of y_t. The observations in y_t also depend on explanatory variables $x_t^k = (x_{1,t}^k \dots x_{S,t}^k)'$, $k = 1, \dots, K$, which are vector transformations of the original univariate explanatory variables. Some of the explanatory variables are constant across equations $s = 1, \dots, S$, $x_{1,t}^k = \dots = x_{S,t}^k$, changing values only with t, while others have distinct values for $s \neq s'$ for the same t. In the case of hourly data the former explanatory variables depend only on the day, whereas the latter depend on the day and on the hour of the day.

Seasonal components with periods longer than S are modelled by regression effects. Each explanatory variable x_t^k is associated with an $S \times 1$ regression coefficient vector $\beta_t^k =$

$(\beta_{1,t}^k \dots \beta_{S,t}^k)'$, $k = 1, \dots, K$. The model is written in matrix form using block diagonal $S \times S$ matrices defined as $B_t^k = \text{diag}(\beta_t^k)$, $k = 1, \dots, K$. Finally, $\varepsilon_t = (\varepsilon_{1,t} \dots \varepsilon_{S,t})'$ is the $S \times 1$ irregular term, a zero mean Gaussian white noise variable with covariance matrix Σ_ε.

The trend μ_t and all regression coefficients β_t^k, $k = 1, \dots, K$, are possibly stochastic and time-varying, following component-specific models: μ_t and β_t^k depend on dynamic factors. The measurement equations are given as:

$$\left\{ \begin{array}{lcl} \mu_t &=& c^0 + \Lambda^0 f_t^0, \\ \beta_t^k &=& c^k + \Lambda^k f_t^k, \end{array} \right. \qquad k{=}1,\dots,\text{K}, \quad t{=}1,\dots,\text{T}, \tag{3.2}$$

with $S \times 1$ vectors $c^j = (c_1^j \dots c_S^j)'$ of constant terms, constant $S \times R^j$ factor loading matrices Λ^j, and R^j dynamic factors in the vectors $f_t^j = (f_{1,t}^j \dots f_{R^j,t}^j)'$, $j = 0, \dots, K$. The number of dynamic factors R^j is specific to each component, with $0 \leq R^j \leq S$. A real factor structure requires $0 < R^j < S$. The constant parameter model is obtained for $R^j = 0$, $j = 0, \dots, K$ and the most general unrestricted model obtains for $R^j = S$, $j = 0, \dots, K$.

The model is completed by the dynamic specification for each vector of dynamic factors f_t^j, $j = 0, \dots, K$. We use two different models, the factors of the trend component follow a local linear trend and the factors of the regression coefficients follow a random walk process.

Formally, we adopt the local linear trend model for $j = 0$. It is given by:

$$\left\{ \begin{array}{lcl} f_{t+1}^j &=& f_t^j + g_t^j + v_t^j, \qquad v_t \sim IIN(0, \Sigma_v^j), \\ g_{t+1}^j &=& g_t^j + w_t^j, \qquad w_t \sim IIN(0, \Sigma_w^j), \end{array} \right. \quad j{=}0, \quad t{=}1,\dots,\text{T}, \tag{3.3}$$

where the vector of dynamic factors f_t^j follows an $R^j \times 1$ multivariate integrated random walk process with a random walk slope vector g_t^j. The vectors v_t^j and w_t^j are $R^j \times 1$ zero mean Gaussian white noise disturbances with covariance matrices Σ_v^j and Σ_w^j. Because f_t^j and g_t^j are clearly nonstationary, f_1^j and g_1^j are initialized with a diffuse distribution as discussed below in section 3.2.6 on state space methods.

We specify a simple random-walk model for the regression coefficient components $j = 1, \dots, K$. This model imposes the restrictions $g_t^j = 0$, $t = 1, \dots, T$, in (3.3) resulting in

$$f_{t+1}^j = f_t^j + e_t^j, \quad e_t^j \sim IIN(0, \Sigma_e^j), \quad j = 1, \dots, K, \quad t = 1, \dots, T, \tag{3.4}$$

where we introduce e_t^j as the $R^j \times 1$ zero mean Gaussian white noise disturbance vector for the regression coefficients of the j-th explanatory variable.

Equations (3.1)-(3.2)-(3.3)-(3.4) form our general multivariate time-varying regression model for periodic time series. For identification purposes and for practical reasons, we impose the following additional restrictions on parameter vectors and matrices:

- *Observation noise covariance matrix*: Although Σ_ε in equation (3.1) can be a general positive semi-definite symmetrical matrix, we prefer a more parsimonious specification in the applications presented below. Therefore we assume throughout that Σ_ε is a non-negative diagonal matrix $\Sigma_\varepsilon = \text{diag}((\sigma^2_{\varepsilon,s})_{s=1,\ldots,S})$.

- *Factor loading matrices*: The $S \times R^j$ factor loading matrices, Λ^j, $j = 0, \ldots, K$ in equation (3.2) are subject to identification restrictions in relation with the corresponding covariance matrices.

- *Constant terms*: We also impose restrictions on the constant terms c^j, $j = 0, \ldots, K$ in equation (3.2) for identification purposes.

- *Covariance matrices of factor disturbances*: The covariance matrices Σ^j_v and Σ^j_w in equation (3.3) and Σ^j_e in equation (3.4) are restricted to be symmetrical and positive semi-definite.

Engle & Watson (1981) and Harvey (1989) (section 8.5) discuss parameter restrictions for identification of dynamic factor models in detail. Here we impose restrictions so that the dynamic factor for each component j, $j = 0, \ldots, K$, corresponds to a subset of R^j elements in the vector y_t. These elements form the basis of our analysis and should therefore have distinct non-zero factor loadings for all components. The corresponding rows in the matrix Λ^j are unit vectors and the corresponding constant terms in c^j are set to 0.

We are also interested in submodels of (3.1)-(3.2)-(3.3)-(3.4) defined by overidentifying restrictions. We distinguish factor restrictions and independence restrictions. Dynamic factors imply reduced dimensions of factor loading matrices in (3.2) and reduced ranks of covariance matrices in (3.3)-(3.4). We follow the literature in the specification of the trend factor f^0_t and we provide a natural extension for the regression coefficient factors f^k_t, $k = 1, \ldots, K$. We also consider independence of the different elements of y_t. Independence implies orthogonality restrictions for all the error terms in (3.1), (3.2), (3.3) and (3.4). In the remainder of this section we discuss the factor specifications and the independent specifications in turn. After a short discussion of the constant parameter model, we conclude this section with a consideration of state space methods for estimation of constant parameters and signal extraction for the time-varying parameters.

3.2.1 General periodic dynamic regression model

The general dynamic regression model corresponds to equations (3.1)-(3.2)-(3.3)-(3.4) in the extreme case where there is no rank reduction in the the dynamic structure and all dynamic factors f_t^j for the components j, $j = 0, \ldots, K$, have dimension $R^j = S$. In this case the factor loading matrices Λ^j in (3.2) reduce to square identity matrices and the constant vectors c^j reduce to zero. The trend component μ_t is a multivariate local linear trend of dimension S with level vector f_t^j and slope vector g_t^j and full rank covariance matrices Σ_v^j, Σ_w^j in equation (3.3). The regression coefficients $j = 1, \ldots, K$, f_t^j follow multivariate random-walks of dimension S with full rank covariance matrices Σ_e^j in equation (3.4)

Model parameters for the periodic dynamic regression model are reduced to matrices Σ_ε, Σ_v^0, Σ_w^0, Σ_e^k, $k = 1, \ldots, K$. Σ_ε is restricted to a positive diagonal matrix. To remain general, there are no a priori constraints on the structure of the dynamics that compose vector y_t. Covariance matrices Σ_v^0, Σ_w^0, Σ_e^k, $k = 1, \ldots, K$ can therefore be full rank : when S is big, it may be difficult to get a consistent estimate of these matrices. Moreover, when some elements of the disturbances vectors are highly correlated, the number of dynamics in the model can be reduced by using dynamic factors in the time-varying regression model framework.

The general dynamic periodic model is implemented for hourly electricity loads in Dordonnat et al. (2008), see also chapter 2. Only bivariate models are implemented and most of the covariance matrices face correlations close to one, motivating the use of dynamic factors for hourly electricity demand modelling and forecasting.

3.2.2 Dynamic factor regression model

The motivation of introducing dynamic factors in the multivariate time-varying regression model is to find more parsimonious specification for periodic time series when some dynamic components disturbances are perfectly correlated. Instead of estimating possibly large covariance matrices, dynamic factors covariance matrices have reduced dimension, and factor loadings together with constant terms have to be estimated.

The dynamic factor regression model corresponds to model (3.1)-(3.2)-(3.3)-(3.4) with reduced dimension for all dynamic factors : $\forall j$, $j = 0, \ldots, K$, $R^j < S$. The number of factors is specific to each stochastic component j. Equation (3.1) is not affected by this restriction. In equation (3.2), factor loading matrices become $S \times R^j$ and for the model to be identified, identification restrictions described earlier are imposed on matrices Λ^j and constant vector c^j,$j = 0, \ldots, K$. In equation (3.3), covariance matrices Σ_v^0, Σ_w^0 of

the stochastic trend component have reduced dimension $R^0 \times R^0$ and in equation (3.4) covariance matrices of the stochastic regression coefficients Σ_e^k, $k = 1, \ldots, K$ reduced dimension $R^k \times R^k$. Moreover, all dynamic factors are supposed independent so that covariance matrices are diagonal ones:

$$\Sigma_v^0 = \text{diag}((\sigma_{v,r}^2)_{r=1,..,R^0}), \quad \Sigma_w^0 = \text{diag}((\sigma_{w,r}^2)_{r=1,..,R^0}),$$
$$\Sigma_e^k = \text{diag}((\sigma_{k,r}^2)_{r=1,..,R^k}), \quad k = 1, \ldots, K$$

Remark : Harvey & Koopman (1997) discuss the specification of common trends, seasonals and cycles for multivariate time series. They illustrate the methodology on several empirical datasets.

3.2.3 Dynamic single factor regression model

The dynamic single factor regression model is the restriction of the dynamic factor regression model to a single dynamic factor for each component. Independent dynamic single factor regression models can be grouped giving a general dynamic factor regression model with block diagonal factor loading matrices.

The model is therefore based on equations (3.1)-(3.2)-(3.3)-(3.4) where all stochastic components (trend and regression coefficients) depend on only one specific dynamic factor, i.e. $R^j = 1$, $j = 0, \ldots, K$. Equation (3.1) remains unchanged. In equation (3.2), factor loading matrices Λ^j, $j = 0, \ldots, K$ are reduced to $S \times 1$ vectors denoted λ^j,$j = 0, \ldots, K$. For the model to be identified, one element of vector y_t is associated with the common dynamic factor (not necessary the same for each component). The corresponding element in λ^j is therefore 1 and the associated constant element in c^j is 0. In equation (3.3), the local linear trend f_t^0 with slope g_t^0 for the trend component μ_t is consequently univariate as well as f_t^k, $k = 1, \ldots, K$ in equation (3.4) become univariate random-walks for all regression coefficients. The covariance matrices to estimate are therefore reduced to scalars σ_v, σ_w and σ_k, $k = 1, \ldots, K$.

3.2.4 Independent univariate time-varying regression model

The independent univariate time-varying regression model reaches parsimony by imposing independence of all stochastic components for all $y_{s,t}$, whereas the general periodic dynamic regression model in section 3.2.1 and the dynamic factor models in sections 3.2.2 and 3.2.3 allow the different elements of vector y_t to be correlated via the underlying dynamics of each component.

Equation (3.1) remains unchanged in the independent specification. As for the general

periodic dynamic regression model, there are as many dynamic factors as the dimension of the vector y_t, so that, imposing the identification restrictions, the matrices Λ^j, $j = 0, \ldots, K$ are identity matrices of dimension S and the constant $S \times 1$ vectors c^j, $j = 0, \ldots, K$ are zero. Equation (3.2) becomes irrelevant. Since we consider the independent univariate case, all covariance matrices in (3.3) and (3.4) Σ_v^0, Σ_w^0 (for the trend component μ_t), Σ_e^k, $k = 1, \ldots, K$ (for all regression coefficients) have to be positive diagonal:

$$\begin{aligned} &\Sigma_v^0 = \text{diag}((\sigma_{v,s}^2)_{s=1,..,S}), \quad \Sigma_w^0 = \text{diag}((\sigma_{w,s}^2)_{s=1,..,S}), \\ &\Sigma_e^k = \text{diag}((\sigma_{k,s}^2)_{s=1,..,S}), \quad k = 1, \ldots, K. \end{aligned}$$

The multivariate model can therefore be estimated independently for the individual equations s=1,..,S, with parameters $\sigma_{\varepsilon,s}, \sigma_{v,s}, \sigma_{w,s}, \sigma_{k,s}$, $k = 1, \ldots, K$. This model will be used as a benchmark model in the next sections.

3.2.5 Constant parameter regression model

All the previous submodels have stochastic components. The constant parameter regression model imposes a deterministic trend $\mu_{t+1} = a + bt$ with $S \times 1$ vectors a and b, and S constant $K \times 1$ regression coefficient vectors β^k, $k = 1, \ldots, K$, form constant matrices B^k. Equations (3.2)-(3.3)-(3.4) become irrelevant so that matrix Σ_ε remains the only hyperparameter matrix to estimate.

The resulting model is a constant parameter periodic regression model. We assume that Σ_ε is diagonal so that each model for $y_{s,t}$ can be estimated independently. If Σ_ε is non-diagonal, the periodic constant regression model is known as a SUR (Seemingly Unrelated Regression) model and can be estimated using generalized least squares.

3.2.6 State-space framework and parameter estimation

This subsection considers essential state space methods. The methods are required for the estimation of the constant unknown (hyper-) parameters and for the signal extraction of the time-varying components.

State-Space Framework

We adopt the general linear Gaussian multivariate state-space model in the following notation of Durbin & Koopman (2001):

$$\left\{ \begin{array}{lcll} y_t & = & Z_t\alpha_t + \varepsilon_t, & \varepsilon_t \sim IIN\left(0, \Sigma_\varepsilon\right), \\ \alpha_{t+1} & = & T_t\alpha_t + R_t\eta_t, & \eta_t \sim IIN\left(0, \Sigma_\eta\right), \end{array} \right. \quad t=1,\ldots,\text{T}. \tag{3.5}$$

The first equation of (3.5) is the measurement (or observation) equation, relating the stochastic vector of observations y_t, the unobserved stochastic state vector α_t and a white noise observation error term. Z_t is the non stochastic, but probably time-varying measurement matrix of appropriate dimensions and ε_t is the observation disturbance (or "noise") term.

The second equation of (3.5) is the transition equation. T_t is the, usually sparse, deterministic transition matrix, R_t is the state disturbance loading matrix and η_t is the white noise state disturbance term affecting α_t. The initial state is given by $\alpha_1 \sim N(a_1, P_1)$.

Textbooks such as Harvey (1989) and Durbin & Koopman (2001) detail model specification, estimation procedures, inference and applications like signal extraction and forecasting for this statistical framework. Widely used linear models and methods fit in this general framework:

- Exponential smoothing methods ,
- Box-Jenkins ARIMA models ,
- Structural (Unobserved Components) Time Series Models (STSM) as introduced by Harvey (1989) and
- Constant and time-varying linear regression models.

Our model combines the latter two models in a multivariate context. State-space models embed stationary and non-stationary components and rely on recursive estimation of the state vector α_t for modelling and forecasting y_t using Kalman filtering and smoothing algorithms. A very relevant feature of the state-space framework for our application is the easy treatment of missing observations for estimation or forecasting: the update of the conditional estimate of the mean and variance of the state vector α_t for the missing observations is trivial. An intricate part of the model specification is the initialization of the state vector α_1. For non-stationary elements of α_t, we suppose the corresponding elements in α_1 are Gaussian with mean $a_1 = 0$ and infinite covariance matrix P_1. Computing recursive estimates with an arbitrary large covariance matrix can lead to significant approximation errors and numerical instability. Durbin & Koopman (2001) present an exact treatment for non-stationary dynamic components and regression coefficients, called diffuse initialization. We use their method to estimate our model and estimate components.

Model (3.1)-(3.2)-(3.3)-(3.4) presented in section 3.2 can easily be written in the state-space framework. Details are given in the appendix for the dynamic single factor model

described in section 3.2.3. Next we describe the estimation procedure for unknown fixed parameters and software implementation of state-space models.

Parameter estimation and signal extraction

Given that all fixed parameters of the system in the measurement matrix Z_t, in the transition matrix T_t, in the error factor loading matrix R_t and in the error covariance matrices Σ_ε and Σ_η are known, the well-known Kalman filtering algorithm can be used to construct the loglikelihood. Indeed, Kalman filtering computes one-step-ahead conditional means and conditional variances for each $y_t, t = 1, \ldots, T$ using the past $\{1, \ldots, t-1\}$. The (diffuse) loglikelihood function of our data is then calculated with the prediction error decomposition.

We estimate the unknown parameters by maximizing the loglikelihood in two steps:

- The Expectation Maximization (EM) algorithm as in Shumway & Stoffer (1982) is applied first, followed by
- Quasi-Newton maximization using numerical scores as in Koopman & Shephard (1992) and Koopman et al. (1999).

The vector of unknown hyperparameters for the general model (3.1)-(3.2)-(3.3)-(3.4) is:

$$\psi = \Big((\sigma_{\varepsilon,s})_{s=1,\ldots,S}, (\sigma_{v,r})_{r=1,..,R^0}, (\sigma_{w,r})_{r=1,..,R^0}, (\sigma_{k,r})_{r=1,..,R^k,k=1,\ldots,K}, \\ vec(\Lambda^0), vec(\Lambda^1), \ldots, vec(\Lambda^K), c^0 \Big)$$

where $vec(M)$ is the vector of all stacked elements in matrix M. The constant vectors c^k, $k = 1, \ldots, K$, are estimated recursively by including these terms in the state vector α_t, as explained in the appendix.

Given the hyperparameter estimates, the Kalman smoothing algorithm can be used to do signal extraction for the time-varying parameters. In our case, the Kalman smoother provides moment estimates of μ_t, B_t^k and $B_t^k x_t^k, k = 1, \ldots, K$, conditional on the whole sample. The Kalman filtering algorithm can also be used to compute k-step ahead forecasts. Forecasting is equivalent to conditional moment (mean and variance) estimation for missing data at the end of the sample. The Kalman filter algorithm also provides standard error estimates, so that confidence intervals are easily derived.

Practical implementation

In the next sections, we present a simulation study and an empirical application based on the models described above. Our modelling strategy is implemented using `Ox`, see Doornik (2006), an object-oriented matrix programming environment. More precisely, the `SsfPack` package of Koopman et al. (1999) and Koopman et al. (2008), provides state-space routines such as Kalman filtering, smoothing and likelihood evaluation for multivariate state space models. Exact treatment of diffuse initial conditions is included. We use these routines to estimate the unknown hyperparameters of our models, for signal extraction and forecasting.

3.3 Monte Carlo experiment

The periodic time-varying regression model with dynamic factors has been presented in section 3.2 and estimation procedures as well as practical implementation have been detailed. This section illustrates the feasibility and interest of our model and methods in a specific Monte Carlo study using typical temperature data as explanatory variables.

3.3.1 Design and plan

For this simulation study, we consider model (3.1)-(3.2)-(3.3)-(3.4) where the seasonal frequency S is equal to 3. For simplicity, we only use one explanatory variable so that $K = 1$. For both the stochastic trend and the regression component we have a dynamic single factor, i.e $R^0 = R^1 = 1$. The length of the multivariate times series is $T = 500$.

The explanatory variable is derived from typical empirical data used in electricity load modelling: it represents heating degrees derived from relevant national temperature averages constructed for France. The series starts in September 1^{st},1995. Heating degrees variables are commonly defined with a threshold temperature, we take a threshold of 15 ^oC

$$x_{s,t} = \max(0, 15 - T_{s,t}), \quad s = 1, \ldots, S, \quad t = 1, \ldots, T,$$

for the temperature on hour s on day t. With this definition, it is supposed that below 15 ^oC heaters start running with an increasing effect on the series $y_{s,t}$ as the temperature goes down. We pick T large enough so that the French National temperature is varying sufficiently below and above the threshold in the simulated datasets. When $x_{s,t} = 0$, $y_{s,t}$ is only determined by a stochastic trend and the stochastic regression coefficient is not identified during these non-heating periods. When $x_{s,t} \neq 0$, $y_{s,t}$ is the sum of a stochastic trend and a stochastic heating effect.

The Data Generating Process (DGP) is specified by equations (3.1)-(3.2)-(3.3)-(3.4). We generate N replications of T observations of the $S \times 1$ observation disturbance vector $\varepsilon_t^{(n)}$ and independent replications for the scalar dynamic component disturbances $v_t^{(n)}, w_t^{(n)}$ and $e_t^{(n)}$. Monte Carlo replications of the time series, $y_t^{(n)}, n = 1, \ldots, N$, are then computed with the constant parameters $\sigma_{\varepsilon,1}, \sigma_{\varepsilon,2}, \sigma_{\varepsilon,3}$ in equation (3.1), $c^0, c^1, \lambda^0, \lambda^1$ in equation (3.2), $\sigma_{v,0}, \sigma_{w,0}$ in equation (3.3) and $\sigma_{e,1}$ in equation (3.4). To satisfy our identification restrictions we impose:

$$\lambda^0 = \begin{pmatrix} 1 & \lambda_2^0 & \lambda_3^0 \end{pmatrix}' \quad , \quad \lambda^1 = \begin{pmatrix} 1 & \lambda_2^1 & \lambda_3^1 \end{pmatrix}'$$
$$c^0 = \begin{pmatrix} 0 & c_2^0 & c_3^0 \end{pmatrix}' \quad , \quad c^1 = \begin{pmatrix} 0 & c_2^1 & c_3^1 \end{pmatrix}'$$

Numerical values for the model parameters are given in Table 3.1. For an easier interpretation of the results, we restricted ourselves to the case $\sigma_{\varepsilon,1} = \sigma_{\varepsilon,2} = \sigma_{\varepsilon,3}$.

We performed $N = 1000$ simulations and for each simulation we followed the estimation procedure described in section 3.2.6, where we estimated two models, A and B:

1. Model A corresponds to the DGP. The unknown parameters estimated by maximizing the loglikelihood function are therefore

$$\left\{\sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3}, \lambda_2^0, \lambda_3^0, c_2^0, c_3^0, \lambda_2^1, \lambda_3^1, \sigma_{v,0}, \sigma_{w,0}, \sigma_{e,1}\right\}.$$

   The unknown parameters c_2^1 and c_3^1 are included in the state vector and consequently recursively estimated by the Kalman filter.

2. Model B is the corresponding multivariate independent model for y_t as presented in section 3.2.4 with S independent dynamic factors for the trend and for the regression component. Separate univariate models for $y_{s,t}$ are estimated. The unknown parameters estimated by maximizing the loglikelihood function are therefore

$$\{\sigma_{\varepsilon,1}, \sigma_{\varepsilon,2}, \sigma_{\varepsilon,3}, \sigma_{v,1}, \sigma_{v,2}, \sigma_{v,3}, \sigma_{w,1}, \sigma_{w,2}, \sigma_{w,3}, \sigma_{e,1}, \sigma_{e,2}, \sigma_{e,3}\}.$$

The expected "pseudo-true" theoretical values for the parameters in model B are based on the DGP parameters and also given in Table 3.1.

Models A and B have the same number of fixed unknown hyperparameters. However, the equations in the misspecified benchmark model B are independent so that we effectively estimate S univariate models, while the DGP involves dependent equations. We investigate the impact of this misspecification on the signal extraction of the state vector in our Monte Carlo analysis.

3.3.2 Monte Carlo Results

We implement the Monte-Carlo study described above in 3.3.1. For each replication (n) we save parameter estimates as well as the smoothed estimates of the stochastic trend and time-varying regression coefficients. We compare the results for models A and B for parameter estimation and signal extraction. Likelihood maximization either for model A (dynamic factor model) or model B (dynamic univariate models) failed in 55 replications. We exclude these replications from the tables and graphs presented below. The effective number of replications is therefore reduced to $N = 945$.

Distribution of the parameter estimates

For each model and each parameter, Table 3.1 presents the true value, the Monte Carlo mean and the Monte Carlo standard deviations of the corresponding estimator.

Table 3.1: Monte Carlo Results for Factor and Univariate models estimation
$N = 945$ Monte-Carlo replications for ML parameter estimates of Model A and Model B on data generated by Model A, as in subsection 3.3.1. Par.: Parameter, True: true value, Mean, s.d: (Monte-Carlo) standard-deviation. Left panel: dynamic factor model A. Right panel: univariate models B. c_2^1 and c_3^1 are elements of the state vector, with estimates for $t = T = 500$.

Factor model (A)				Univariate models (B)			
Par.	True	Mean	S.d.	Par.	True	Mean	s.d.
$\sigma_{\varepsilon,1}$	200	199.0	6.5	$\sigma_{\varepsilon,1}$	200	198.8	7.1
$\sigma_{\varepsilon,2}$	200	199.3	7.3	$\sigma_{\varepsilon,1}$	200	199.3	7.7
$\sigma_{\varepsilon,3}$	200	199.3	6.3	$\sigma_{\varepsilon,1}$	200	199.1	6.6
$\sigma_{v,0}$	6	7.1	8.2	$\sigma_{v,1} = \sigma_{v,0}$	6	11.2	13.1
$\sigma_{w,0}$	3	2.8	0.6	$\sigma_{v,2} = \lambda_2^0 \sigma_{v,0}$	12	15.6	17.2
$\sigma_{e,1}$	5	4.9	0.8	$\sigma_{v,3} = \lambda_3^0 \sigma_{v,0}$	3	7.9	9.5
λ_2^0	2	2.0	0.01	$\sigma_{w,1} = \sigma_{w,0}$	3	2.7	0.7
λ_3^0	0.5	0.5	0.01	$\sigma_{w,2} = \lambda_2^0 \sigma_{w,0}$	6	5.6	1.2
λ_2^1	2	2.0	0.1	$\sigma_{w,3} = \lambda_3^0 \sigma_{w,0}$	1.5	1.3	0.4
λ_3^1	0.5	0.5	0.1	$\sigma_{e,1} = \sigma_{e,1}$	5	4.7	1.4
c_2^0	2	2.7	40.9	$\sigma_{e,2} = \lambda_2^1 \sigma_{e,1}$	10	9.7	1.7
c_3^0	4	3.9	20.6	$\sigma_{e,3} = \lambda_3^1 \sigma_{e,1}$	2.5	2.1	1.1
c_2^1	5	-2.0	73.0				
c_3^1	10	7.7	36.3				

We first focus on parameter estimation for model A, which is the model used to simulate data. The irregular standard deviation estimates are nearly unbiased. For $\sigma_{\varepsilon,s}$, $s = 1, 2, 3$, the empirical means are very close to the true value of 200 for model A, also for model B. Their standard deviations are relatively small. Model A contains three other standard deviation parameters: $\sigma_{v,0}, \sigma_{w,0}$ for the trend, and $\sigma_{e,1}$ for the regression component. The estimate for the level component is a little biased : the true value is

6 and the empirical mean of the estimator is 7.1. Not also that its standard-deviation is large relative to the true value (8.2). In comparison, the means are closer to the true values for the slope and the regression component, and the standard deviations of these estimates are much smaller. The factor loadings for model A are very precisely estimated, the means of the estimators are very close to the true value and the standard deviations are very small, both for the trend loadings (λ_s^0, $s = 2, 3$) and for the regression coefficient loadings $\lambda_s^1, s = 2, 3$. In contrast, the estimation of the constant terms c_s^0, $s = 2, 3$ for the trend and c_s^1, $s = 2, 3$ for the regression effects are somewhat unprecise, with a bias and relatively large standard deviations.

Model B contains six standard deviation parameters $\sigma_{v,s}, \sigma_{w,s}$, $s = 1, 2, 3$ for the trend and three $\sigma_{e,s}$, $s = 1, 2, 3$ for the regression effect. This model can only estimate "pseudo-true" values, but we compare the estimators with the corresponding parameters in model A. Again, there is a clear bias in the estimators for the level component parameters, with large standard deviations. Results are more satisfactory for the estimation of the variation in the slope and in the regression coefficients.

For a more complete picture, we also present graphical output of the empirical distribution of the parameter estimator. We present the histogram and nonparametric density estimate (black, continuous line) of the estimator of each coefficient. We also show the Gaussian approximation (blue, dotted line), i.e. the normal distribution with the same expectation and standard-deviation as the estimator.

Figure 3.1 shows the standard-deviation estimates for model A in panels (a): $\hat{\sigma}_{w,0}$, (b): $\hat{\sigma}_{v,0}$ and (c): $\hat{\sigma}_{e,1}$. Figure 3.1(a) shows that the empirical density for the estimates of the slope component standard deviation is well approximated by the corresponding Gaussian density. The true value for this standard deviation is $\sigma_{w,0} = 3$. Figure 3.1(b) shows that the empirical density for the estimates of the level component standard deviation $\sigma_{v,0}$ is flat, compared to a peak near zero. This peak indicates a discrete component of the distribution. In the literature, this feature of variance estimators in unobserved component models is known as the pile-up problem, see e.g. Shephard (1993) or Stock & Watson (1998) for more details. It is due to the fact that the estimate is constrained to be strictly positive and that the true value is relatively small (here we have $\sigma_{v,0} = 6$). Figure 3.1(c) shows that the empirical density for the estimates of the regression effect standard deviation $\sigma_{e,1}$ is well adjusted by the corresponding Gaussian density, similar to $\hat{\sigma}_{w,0}$. Figure 3.1 also displays the empirical distribution and Gaussian approximations for the estimates of the measurement equation standard deviations $\hat{\sigma}_{\varepsilon,s}$, $s = 1, 2, 3$ in panels (d), (e) and (f). These distributions are very close to Gaussian and there is no clear bias.

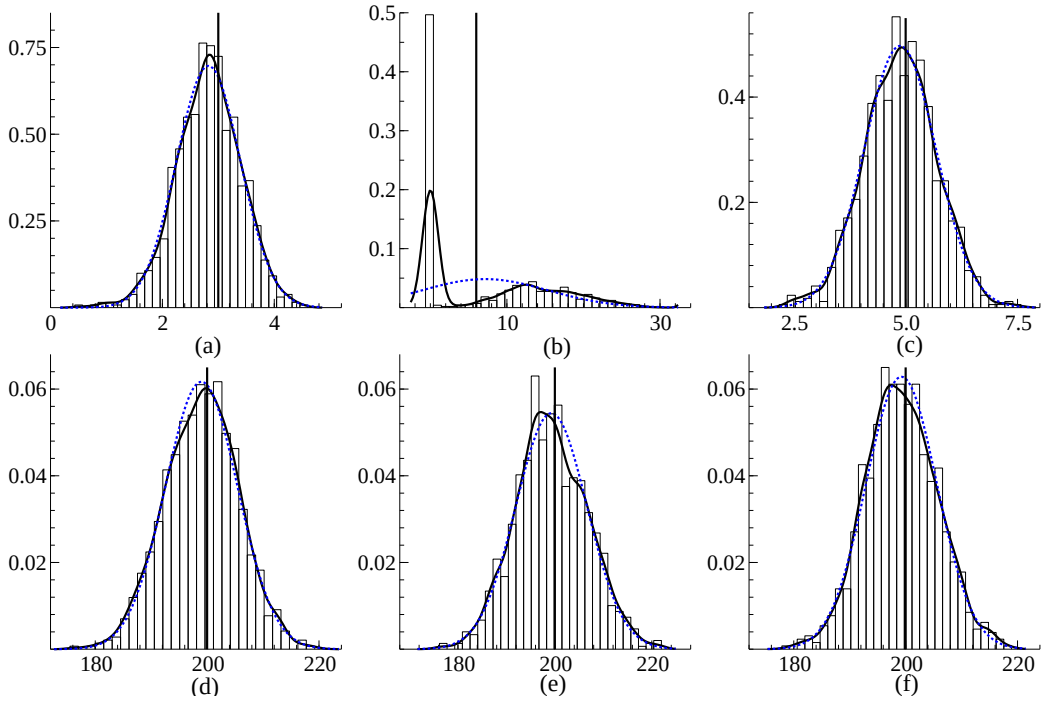

Figure 3.1: Simulation study - Estimation results for model A. Empirical distribution (histogram, density- black continuous line) and normal approximation (blue dotted line) of the estimates for the standard deviations of transition equations (3.3)-(3.4) and observation equation (3.1) : (a)$\hat{\sigma}_{w,0}$, (b) $\hat{\sigma}_{v,0}$, (c) $\hat{\sigma}_{e,1}$,(d)$\hat{\sigma}_{\varepsilon,1}$, (e) $\hat{\sigma}_{\varepsilon,2}$, (f) $\hat{\sigma}_{\varepsilon,3}$, see also Table 3.1.

Figure 3.2 shows the empirical distribution and Gaussian approximations for the factor loading estimates in panel (a): $\hat{\lambda}_2^0$, (b):$\hat{\lambda}_3^0$, (c): $\hat{\lambda}_2^1$ and (d): $\hat{\lambda}_3^1$. These estimators are not constrained so there can be no pile-up. Empirical densities have their mean close to the true values $\lambda_2^0 = \lambda_2^1 = 2, \lambda_3^0 = \lambda_3^1 = 0.5$, and they are more concentrated around the mean than their Gaussian approximations. Figure 3.2 also shows the distributions for the estimators of the constant terms of the factor loading equations. Panels 3.2 (e), (f), (g) and (h) show $\hat{c}_2^0$, $\hat{c}_3^0$, $\hat{c}_2^1$, $\hat{c}_3^1$, respectively. The coefficients c_2^0 and c_3^0 are directly estimated as hyperparameters during the likelihood maximization step whereas c_2^1 and c_3^1 are estimated recursively in the state vector α_t using Kalman smoother. The distributions of the estimators are close to Gaussian. However, their standard deviations are large.

The results in Figure 3.1 show estimates for the true dynamic factor model using the correct model A. These results can be compared with the estimation results for Model B, i.e. for the corresponding (independent) univariate models. Figure 3.3 displays the empirical distributions and Gaussian approximations for the standard-deviation estimates using model B: Figure 3.3(a) shows $\hat{\sigma}_{w,1}$, the other panels (b): $\hat{\sigma}_{v,1}$, (c): $\hat{\sigma}_{e,1}$, (d): $\hat{\sigma}_{w,2}$, (e): $\hat{\sigma}_{v,2}$, (f): $\hat{\sigma}_{e,2}$, (g): $\hat{\sigma}_{w,3}$, (h): $\hat{\sigma}_{v,3}$, and finally, panel 3.3(i) shows $\hat{\sigma}_{e,3}$. The empirical distributions for $\hat{\sigma}_{w,1}$, $\hat{\sigma}_{w,2}$, $\hat{\sigma}_{w,3}$, shown in panels (a),(d) and (g), are very well

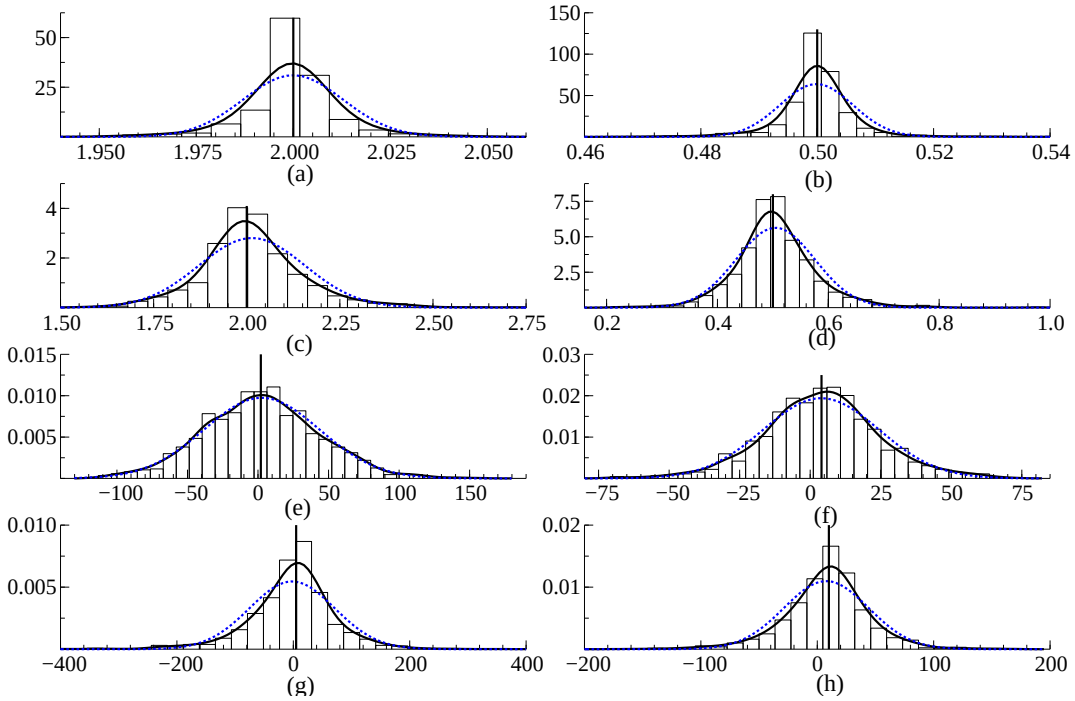

Figure 3.2: Simulation study - Estimation results for model A. Empirical distribution (histogram, density- black continuous line) and normal approximation (blue dotted line) of the estimates for the factor loadings and constant terms of equation (3.2) : (a) $\hat{\lambda}_2^0$, (b) $\hat{\lambda}_3^0$,(c) $\hat{\lambda}_2^1$, (d) $\hat{\lambda}_3^1$, (e) $\hat{c}_2^0$, (f) $\hat{c}_3^0$, (g) $\hat{c}_2^1$, (h) $\hat{c}_3^1$, see also Table 3.1.

approximated a Gaussian distribution, but they are a bit biased. Similar to the results for the factor model A, the univariate estimates of $\sigma_{v,1}$, $\sigma_{v,2}$, $\sigma_{v,3}$ in panels (b), (e) and (h) exhibit the pile-up and a bias when comparing the estimates with the "pseudo-true" values $\sigma_{v,1} = 6, \sigma_{v,2} = 12, \sigma_{v,3} = 3$. Regarding the regression effect standard deviation estimates in panels (c), (f) and (i), the Gaussian approximation is good and there is virtually no bias for $\hat{\sigma}_{e,1}$. For $\hat{\sigma}_{e,2}$, the Gaussian approximation is a little worse with a longer left tail for the empirical density. The distribution of $\hat{\sigma}_{e,3}$ also exhibits a clear pile-up phenomenon, and a clear bias. The results for $\hat{\sigma}_{\varepsilon,s}, s = 1, 2, 3$ for the univariate models B are similar to those for the true model A and therefore not shown.

Signal extraction of the state vector

After discussing the accuracy of hyperparameter estimators, we continue the comparison between the performance of the true dynamic factor model A and the univariate benchmark model B by considering signal extraction. For both models and each draw (n), we ran the Kalman smoothing algorithm with estimated hyperparameter values and we obtained smoothed estimates of the trend component and the stochastic regression coefficient, based on the full simulated sample. In a Monte Carlo analysis these smoothed

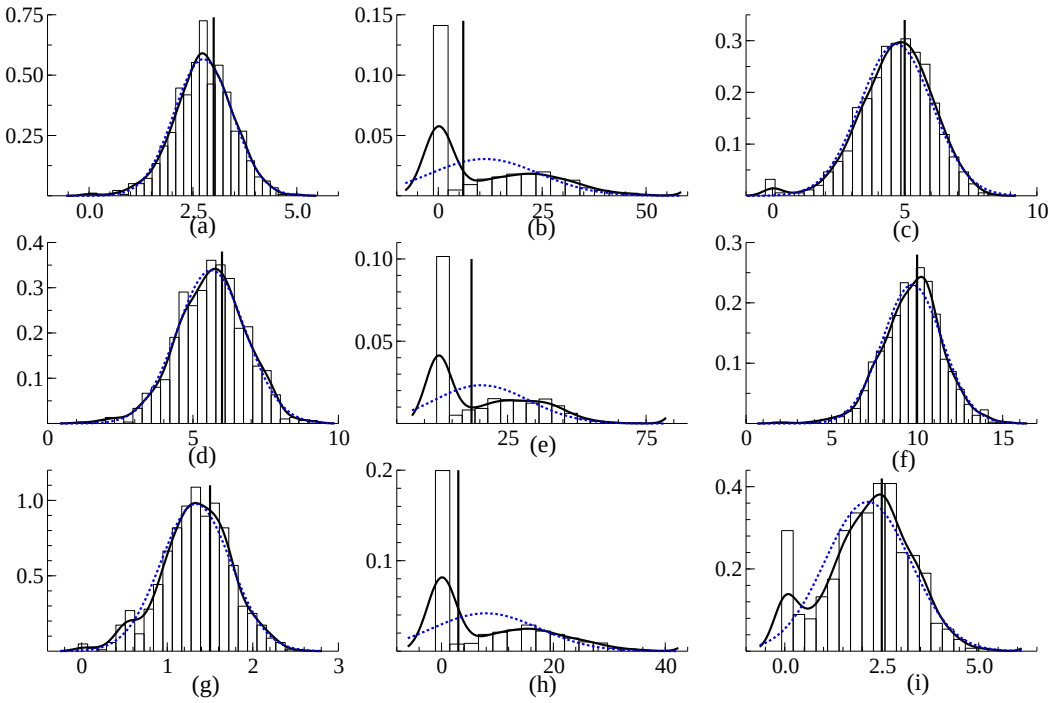

Figure 3.3: Simulation study - Estimation results for model B. Empirical distribution (histogram, density- black continuous line) and normal approximation (blue dotted line) of the estimates for the standard deviations of transition equations (3.3)-(3.4) : (a) $\hat{\sigma}_{w,1}$, (b) $\hat{\sigma}_{v,1}$, (c) $\hat{\sigma}_{e,1}$,(d) $\hat{\sigma}_{w,2}$, (e) $\hat{\sigma}_{v,2}$, (f) $\hat{\sigma}_{e,2}$, (g) $\hat{\sigma}_{w,3}$, (h) $\hat{\sigma}_{v,3}$, (i) $\hat{\sigma}_{e,3}$, see also Table 3.1.

estimates can be compared with the "true" underlying simulated signal in each draw. We measure the accuracy of the stochastic trend and regression coefficient estimation for each time point using the well-known RMSE (Root Mean Squared Error) to measure the accuracy of the stochastic trend and regression coefficient estimation for each point of the sample.

Figure 3.4 shows the simulation results. Figures 3.4 (a), (d) and (g) first show the explanatory variable used for the simulations. We can see two periods where this variable is non-zero. The regression coefficient can therefore only be estimated properly for these periods. Figures 3.4 (b), (e) and (h) show the time-varying RMSE for stochastic trend extraction for model A (purple, bold line) and B (blue, dotted line). The factor model A outperforms the univariate models B, most clearly for $s = 1$ and $s = 3$. We also notice that signal extraction is much better for both models when the explanatory variable is zero. Figures 3.4 (c), (f) and (i) show the signal extraction accuracy for the states of the time-varying-regression coefficient for factor model A and univariate model B. The value of the RMSE is set to zero when the explanatory variable is zero. As for the trend, model A outperforms model B, most clearly for $s = 1$ and $s = 3$.

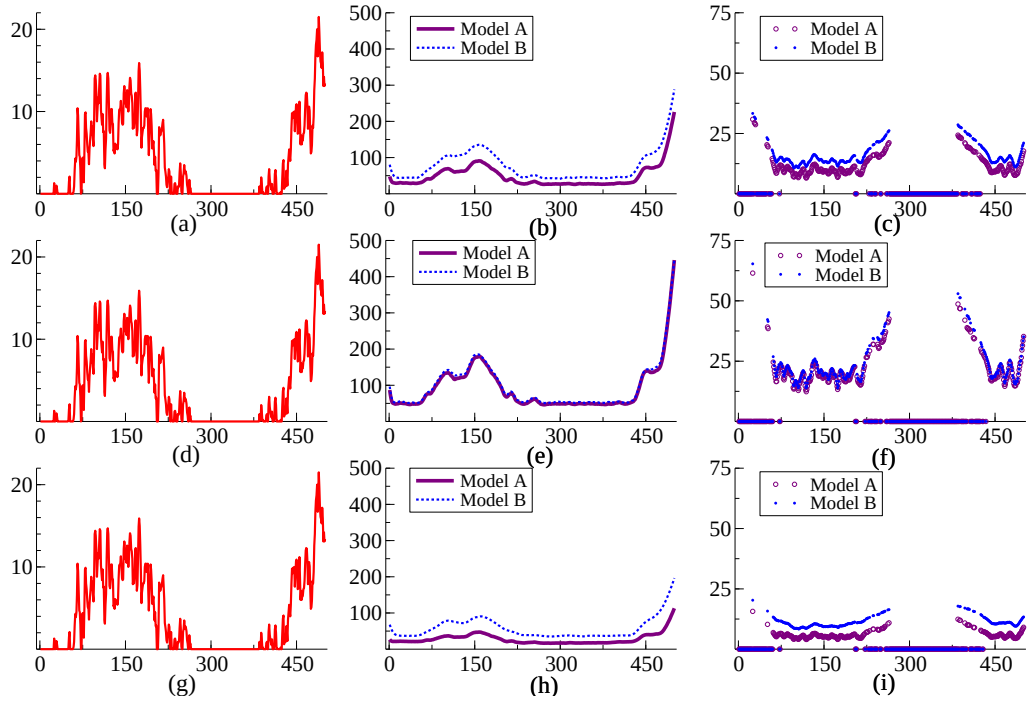

Figure 3.4: Simulation study - Signal extraction accuracy. Explanatory variable and RMSE for the smoothed estimate of the state vector α_t for factor model A (purple bold line) and univariate model B (blue dotted line) : (a)-(d)-(g) Explanatory variable $x_{s,t}, s = 1, 2, 3$, (b)-(e)-(h) RMSE for the smoothed estimate of $\mu_{s,t}, s = 1, 2, 3$, (c)-(f)-(i) RMSE for the smoothed estimate of $\beta^1_{s,t}, s = 1, 2, 3$; $t = 1, \ldots, T = 500$.

3.3.3 Summary of Monte Carlo results

Our specific simulation study compared the performance of the dynamic factor model and the corresponding univariate models for a simple dynamic single factor DGP of a single stochastic trend plus a one-factor stochastic dynamic regression effect for a trivariate time series. We focussed on the distribution of the parameter estimates and on signal extraction accuracy. While the effects of the misspecification of the benchmark model are slightly ambiguous for hyperparameter estimation, it is clear that the use of separate univariate models when the time series are really related via dynamic factors leads to an important loss in the accuracy of signal extraction, in comparison with using the true factor model.

3.4 Empirical modelling of national French hourly electricity loads

The methodology described in section 3.2 is applied to model and forecast hourly electricity loads in France. In this application, we therefore consider $S = 24$. To obtain

useful results in practice, we split the $S = 24$ series into subgroups with separate dynamic factors for the trends and regression coefficients. We first describe the dataset. Next, we detail the full model for hourly loads. Finally, we present estimation results, signal extraction and we examine the short-term forecasting accuracy of our model.

3.4.1 Data description

The time series of hourly electricity loads that we analyse has many typical interesting features: a long term (positive) trend, different levels of seasonality (yearly, weekly, daily), influence of weather variables such as temperature and cloud cover. We model aggregate hourly electricity loads in France, measured in MegaWatts (MW). The dataset is long enough (from January 1^{st}, 1997 to August 31^{st}, 2004) to study long-term changes in components and effects. This dataset has been previously described and used in Dordonnat et al. (2008), see chapter 2 for more details. French data are affected by special days which alter the seasonal cycle of the series e.g. : Bank holidays, special periods around the end of the year (roughly from December 23^{rd} to January 3^{rd}), daylight saving days and the so-called "EJP" days (during those days, there is a financial incentive to reduce electricity consumption and they therefore require a special treatment). Forecasting electricity loads for those days requires special expertise. These days are therefore excluded from the study to focus on more general patterns in consumption behaviour.

In this application, we do not aim at building the best load forecasting model but we focus on the benefits we can get for signal extraction in comparison with separate independent modelling of each hour.

3.4.2 Empirical model and implementation

The model for French hourly electricity loads fits in the general model (3.1)-(3.2)-(3.3)-(3.4). We model electricity loads by groups of three consecutive hours to remain practical, so that matrices Λ^0, Λ^k, $k = 1, \ldots, K$ in equation (3.2) have block diagonal structures, just like $\Sigma_w^0, \Sigma_v^0, \Sigma_e^k$, $k = 1, \ldots, K$ in equations (3.3)-(3.4). For convenience, we consider $s = 0, \ldots, 23$ instead of $s = 1, \ldots, 24$ for the hour index. We denote by S^F the subset of hours related to common factors (i.e. hours for which the corresponding rows in Λ^0 and each Λ^k are unit vectors and related constant terms in c^0 and each c^k are equal to 0): we choose $S^F = \{0; 3; 6; 9; 12; 15; 18; 21\}$. We denote by S^{NF} the subset of hours which are not in S^F.

Informally, the empirical model decomposes French electricity demand as follows:

$$y_t = \text{Trend}_t + \text{Yearly Seasonal}_t + \text{Weekly Seasonal}_t + \text{WeatherEffects}_t + \text{Extras}_t + \varepsilon_t. \quad (3.6)$$

Next we provide the details of the different components and their formal models.

Trend specification

The model for hourly electricity loads first includes a long-term trend. The $S \times 1$ vector μ_t contains the trend for each hour. Since we estimate 8 independent trivariate models, matrix Λ^0 in (3.2) is block-diagonal and for each block of 3 hours, the specification corresponds to a dynamic single factor. The full specification of the first equation in equation (3.2) is therefore:

$$\Lambda^0 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \lambda^0_{1,1} & \vdots & & \vdots \\ \lambda^0_{1,2} & 0 & \dots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \dots & 0 & 1 \\ \vdots & & \vdots & \lambda^0_{8,22} \\ 0 & \dots & 0 & \lambda^0_{8,23} \end{pmatrix}, \quad c^0 = \begin{pmatrix} 0 \\ c^0_2 \\ c^0_3 \\ \vdots \\ 0 \\ c^0_{22} \\ c^0_{23} \end{pmatrix}, \tag{3.7}$$

where the corresponding f^0_t is an 8×1 vector of common local linear trends, shared by groups of 3 hours, specified as in (3.3). As Σ^0_v and Σ^0_w are assumed diagonal, the model can be separately estimated for each group of 3 hours.

Remark : To impose smooth trends, the variance values for each common trend are set to 0.1 (level and slope). Larger values lead to estimated trends that capture the yearly cycle of the time series which is an undesirable feature. Using these pre-fixed variance values, we obtain a smooth long-term trend. Fixing some parameters value is required to obtain a good decomposition when the data are no informative enough.

Stochastic regression effects

The daily periodicity of hourly electricity demand is captured by the vector structure of the model. The yearly cycle, weekly pattern and weather influence are modelled by time-varying regression effects. Three hours of the day display a different type of weekly pattern and interaction effects of the yearly cycle. We label these as extra effects.

Yearly cycle : The yearly patterns are captured with Fourier terms as regressors:

$$\begin{aligned} &x^1_{s,t} = a_{1,t},\ x^2_{s,t} = b_{1,t},\ x^3_{s,t} = a_{2,t},\ x^4_{s,t} = b_{2,t}, \quad s = 0, \dots, 23, \\ &a_{i,t} = \cos\left(\tau_t \tfrac{2\pi i}{365.25}\right), \quad b_{i,t} = \sin\left(\tau_t \tfrac{2\pi i}{365.25}\right), \quad i = 1, 2, \end{aligned} \tag{3.8}$$

where τ_t is the number of days elapsed since the 1^{st} of January in the year in which day t falls for $t = 1, \ldots, T$, and i is the frequency in cycles per year.

Weekly pattern : The weekly pattern is modelled by four dummy variables for day types:

- $x^5_{s,t} = 1$ if day t is a Monday, 0 otherwise;
- $x^6_{s,t} = 1$ if day t is a Friday, 0 otherwise;
- $x^7_{s,t} = 1$ if day t is a Saturday, 0 otherwise;
- $x^8_{s,t} = 1$ if day t is a Sunday, 0 otherwise.

The default baseline day type for the trend estimates therefore corresponds to Tuesdays, Wednesdays and Thursdays.

Extra effects : To obtain satisfactory diagnostic results for the dynamic specification of the group of three morning hours 6, 7 and 8, we introduce distinct yearly cycles for weekdays and weekends. For these hours we redefine $x^k_{s,t}, k = 1, \ldots, 4$ to equal the Fourier terms as defined in (3.8) on weekdays and to be zero on weekends. Extra regressors $x^k_{s,t}, k = 13, \ldots, 16$, equal the Fourier terms in weekends, being zero on weekdays. We also introduces two extra dummy daytype variables, $x^{17}_{s,t}, x^{18}_{s,t}$, for Wednesdays and Thursdays. The baseline daytype for these morning hours is therefore Tuesday.

Weather variables : The weather variables used in the regression part are the same as in Chapter 2, where we present more details. We distinguish

- $x^9_{s,t} = \max(0, 15 - T_{s,t})$ corresponding to heating degrees based on instantly observed national temperature $T_{s,t}$.
- $x^{10}_{s,t} = \max(0, 15 - T^{smo}_{t,s})$ is a smoothed-heating degrees variable, based on an exponential smoothing of the temperature variable $T^{smo}_{t,s}$,
- $x^{11}_{s,t} = \max(0, T^{smo}_{t,s} - 18)$ is a smoothed-cooling degrees variable $T^{smo}_{t,s}$,
- The last weather variable $x^{12}_{s,t}$ is a measure of cloud cover in heating periods, being non-zero only when $x^9_{s,t} > 0$.

Dynamic specification for the regression coefficients

The model for the stochastic regression coefficients β_t^k for variables $x_t^1, ..., x_t^{10}$, is given by (3.2) and (3.4). For all k, we specify Λ^k and c^k in (3.2) in the same way as in (3.7):

$$\Lambda^k = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \lambda_{1,1}^k & \vdots & & \vdots \\ \lambda_{1,2}^k & 0 & \dots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \dots & 0 & 1 \\ \vdots & & \vdots & \lambda_{8,22}^k \\ 0 & \dots & 0 & \lambda_{8,23}^k \end{pmatrix}, \quad c^k = \begin{pmatrix} 0 \\ c_2^k \\ c_3^k \\ \vdots \\ 0 \\ c_{22}^k \\ c_{23}^k \end{pmatrix}, \quad k = 1, \dots, 10. \tag{3.9}$$

The corresponding 8×1 vectors of factors f_t^k, $k = 1, \dots, 10$, follow random walk processes as in (3.4) with diagonal disturbance covariance matrices Σ_e^k. The specification for the coefficients of the cooling degrees variable x_t^{11} follows section 3.2.4 with independent random walk components for all hours so that Σ_e^{11} is a positive diagonal matrix of dimension 24×24. The effect for the cloud-cover variable x_t^{12} is taken as a constant $\beta_t^{12} = \beta^{12}$ for $t = 1, \dots, T$. Finally, the dynamic specification for the regression coefficients $\beta_t^k, k = 13, \dots, 18$, for the extra regressors for the group of morning hours 6, 7 and 8, is given by (3.2) and (3.4). We specify Λ^k and c^k in (3.2) as follows:

$$\Lambda^k = \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \lambda_{3,7}^k & \vdots & & \vdots \\ 0 & \dots & 0 & \lambda_{3,8}^k & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad c^k = \begin{pmatrix} 0_{6\text{x}1} \\ 0 \\ c_7^k \\ c_8^k \\ 0_{15\text{x}1} \end{pmatrix}, \quad k = 13, \dots, 18. \tag{3.10}$$

A univariate random walk dynamic factor f_t^k as in (3.4) is specified for each $k = 13, \dots, 18$.

This completes the formal specification of the informal model (3.6). In sum, the full model reads,

$$y_t = \mu_t + \sum_{k=1}^{10} B_t^k x_t^k + B_t^{11} x_t^{11} + B^{12} x_t^{12} + \sum_{k=13}^{18} B_t^k x_t^k + \varepsilon_t, \tag{3.11}$$

where components μ_t, $\sum_{k=1}^{8} B_t^k x_t^k$, and $\sum_{k=13}^{18} B_t^k x_t^k$ are defined near (3.7), (3.9) and (3.10), respectively, where the diagonal elements of B_t^{11} follow 24 independent random walks and where the diagonal elements of B_t^{11} are constant.

Benchmark model

In order to examine the practical advantages of dynamic factor modelling of hourly electricity loads, we compare our model with a benchmark of 24 independent dynamic models for each hour as in section 3.2.4. The univariate benchmark models for the different hours of the day are:

$$y_{s,t} = \mu_{s,t} + \sum_{k=1}^{11} \beta_{s,t}^k x_{s,t}^k + \beta_s^{12} x_{s,t}^{12} + I_{[6,7,8]}(s) \cdot \sum_{k=13}^{18} \beta_{s,t}^k x_{s,t}^k + \varepsilon_{s,t}, \quad s = 0, \ldots, 23, t = 1, \ldots, T, \tag{3.12}$$

with $I_{[6,7,8]}$ an indicator function equal to one for $s = 6, 7, 8$ and zero elsewhere and where $\mu_{s,t}$ and $\beta_{s,t}^k$, $k = 1, \ldots, K = 18$, follow the univariate versions of the specifications in (3.11).

Practical implementation

The final model for electricity loads is put in state-space form as explained in section 3.2.6 and is implemented using using the matrix language `Ox` by Doornik (2006) and the `SsfPack` routines by Koopman et al. (2008). The dataset is split into two parts: in-sample corresponds to the period January 1^{st}, 1997 to August 31^{st}, 2003 and post-sample corresponds to the last year of data, September 1^{st}, 2003 to August 31^{st}, 2004. Since the model is built from blocks of independent trivariate models, we can estimate the 8 models successively. For each trivariate model r, $r = 1, \ldots, 8$, the vector of unknown parameters is $(\psi_{\sigma,r}, \psi_{\lambda,r}, \psi_{c,r})$ where $\psi_{\sigma,r}$ is defined as:

$$\psi_{\sigma,r} = \left(\sigma_{\varepsilon,1(r)}, \ldots, \sigma_{\varepsilon,3(r)}, \sigma_{v(r)}, \sigma_{w(r)}, \sigma_{1(r)}, \ldots, \sigma_{10(r)}, \sigma_{11,1(r)}, \ldots, \sigma_{11,3(r)}\right)$$

We perform unconstrained optimisation of the parameters. Since the standard deviation parameters have to be strictly positive, the optimisation is done using the transformation $\ln(\psi_{\sigma,r})$. We further define

$$\psi_{\lambda,r} = \left(\lambda_{2(r)}^0, \lambda_{3(r)}^0, \lambda_{2(r)}^1, \lambda_{3(r)}^1, \ldots, \lambda_{2(r)}^{10}, \lambda_{3(r)}^{10}\right), \quad \psi_{c,r} = \left(c_{2(r)}^0, c_{3(r)}^0, \ldots, c_{2(r)}^{10}, c_{3(r)}^{10}\right),$$

where $\psi_{\lambda,r}, \psi_{c,r}$ are unconstrained parameters, therefore no transformation is required for the likelihood maximization step. For $r = 3$, i.e. for the group of three morning hours 6, 7, and 8, the vector of parameters $\psi_{\sigma,3}$ is extended with $\left(\sigma_{13(3)}, \ldots, \sigma_{18(3)}\right)$, $\psi_{\lambda,3}$ with

$\left(\lambda^{13}_{2(3)}, \lambda^{13}_{3(3)}, \ldots, \lambda^{18}_{2(3)}, \lambda^{18}_{3(3)}\right)$ and $\psi_{c,3}$ with $\left(c^{13}_{2(3)}, c^{13}_{3(3)}, \ldots, c^{18}_{2(3)}, c^{18}_{3(3)}\right)$. Note that in the notation of (3.9) $\lambda^k_{j(r)} = \lambda^k_{r,3(r-1)+j-1}, c^k_{j(r)} = c^k_{3(r-1)+j-1}, k = 0, \ldots, 10, r = 1, \ldots, 8, j = 2, 3$, and further $k = 13, \ldots, 18, r = 3$.

Remark : For the likelihood maximization step, the initial value for each $c^0_{j(r)}, j = 2, 3$ is $\bar{y}_{j(r)} - \bar{y}_{1(r)}$, i.e. the average load difference of hour $j(r)$ with the baseline hour $1(r)$ for the corresponding trend factor. The other constant terms $c^k_{j(r)}, k = 1, \ldots, 10$ (and further $k = 13, \ldots, 18$ for $r = 3$) are recursively estimated by putting them in the state vector, as detailed in the appendix.

For benchmark model (3.12), the vector ψ_s of unknown parameters to estimate for each $s = 0, \ldots, 23$ is:

$$\psi_s = (\sigma_{\varepsilon,s}, \sigma_{v,s}, \sigma_{w,s}, \sigma_{1,s}, \ldots, \sigma_{11,s}),$$

extended for $s = 6, 7, 8$ with $(\sigma_{13,s}, \ldots, \sigma_{18,s})$.

All parameters are standard deviations so that the likelihood maximization is performed with respect to $\ln(\psi_s)$.

3.4.3 Estimation results

Table 3.2: Parameter Estimates $\hat{\sigma}_{k(r)}$ of dynamic factor model for French load I
Parameter estimates for the empirical model of French electricity loads, see section 3.4.2. Sample: January 1, 1997 - August 31, 2003. Estimated standard deviations for the factors related to stochastic regression effects: $\hat{\sigma}_{k(r)}$, $k = 1, .., 10, 13, \ldots, 18$; $r = 1, .., 8$, index r correspond to hours in S^F.
$r = 1$: hour 0; $r = 2$: hour 3 etc. See also Table 3.3 and Table 3.4.

	Component	Par.	r=1	r=2	r=3	r=4	r=5	r=6	r=7	r=8
f^1_t	$a_{1,t}$	$\hat{\sigma}_{1(r)}$	824.1	665.9	311.4	686.5	530.8	429.8	464.5	388.3
f^2_t	$b_{1,t}$	$\hat{\sigma}_{2(r)}$	683.6	720.9	385.0	470.6	518.4	627.3	449.8	258.9
f^3_t	$a_{2,t}$	$\hat{\sigma}_{3(r)}$	24.2	112.1	39.5	388.3	305.7	4.1	614.6	226.2
f^4_t	$b_{2,t}$	$\hat{\sigma}_{4(r)}$	396.9	330.7	480	619.4	336.2	644.6	822.3	448.5
f^5_t	Monday	$\hat{\sigma}_{5(r)}$	3.5	8.4	8.2	43.7	6.9	5.5	77.4	2.3
f^6_t	Friday	$\hat{\sigma}_{6(r)}$	0.7	0.7	3.0	1.1	0.3	34.2	4.5	4.1
f^7_t	Saturday	$\hat{\sigma}_{7(r)}$	2.6	34.8	208.9	68.0	105.7	130.4	77.3	22.1
f^8_t	Sunday	$\hat{\sigma}_{8(r)}$	31.7	26.8	14.7	92.1	103.4	190.5	13.2	129.4
f^9_t	Heating	$\hat{\sigma}_{9(r)}$	91.0	30.4	36.3	24.1	21.7	32.8	21.1	33.8
f^{10}_t	Smoothed-heating	$\hat{\sigma}_{10(r)}$	5.1	19.0	25.0	34.6	78.1	78.0	43.9	1.3
f^{13}_t	$a^{WE}_{1,t}$	$\hat{\sigma}_{13(r)}$			15.2					
f^{14}_t	$b^{WE}_{2,t}$	$\hat{\sigma}_{14(r)}$			41.6					
f^{15}_t	$a^{WE}_{1,t}$	$\hat{\sigma}_{15(r)}$			91.3					
f^{16}_t	$b^{WE}_{2,t}$	$\hat{\sigma}_{16(r)}$			292.9					
f^{17}_t	Wednesday	$\hat{\sigma}_{17(r)}$			33.4					
f^{18}_t	Thursday	$\hat{\sigma}_{18(r)}$			8.7					

Tables 3.2 and 3.3 present parameter estimates after independent likelihood maximization for each trivariate model on the estimation sample January 1, 1997 until August

31, 2003. Table 3.2 presents estimates of standard deviations for the factor dynamics corresponding to each stochastic regression effects for hours s in S^F. Table 3.3 contains estimated standard deviations for the hourly independent cooling effect and the irregular term from observation equation (3.1); it also contains estimates $\hat{\beta}_s^{12}$ for the constant regression effect of the cloud-cover variable with associated t-value in brackets.

For the Fourier coefficients, almost all estimated standard deviations in Table 3.2 are very large, especially when compared with the irregular standard deviations in Table 3.3. The estimated yearly pattern is therefore highly time-varying and may cause forecasting problems. The estimated standard deviations for Mondays and Fridays are relatively small except for the morning hours (9,10,11) and the evening peak hours (18,19,20) for the Monday effect and for the afternoon hours (15,16,17) for the Friday effect: the effect for those days is quite smooth. Daytype effects are highly time-varying for the weekends during the day. For heating and smoothed-heating effects, the heating effect tends to be more time-varying during night hours than the smoothed-heating effect and less time-varying during the day. The estimated standard deviations $\hat{\sigma}_{13(3)}$ to $\hat{\sigma}_{18(3)}$ are more difficult to interpret since they are only estimated for one group of three morning hours (6, 7, 8). For those hours, the Wednesday effect is the second most time-varying day-type effect after the Saturday effect.

Standard deviations for the cooling effect in Table 3.3 are strictly positive and relatively small: this component is time-varying but relatively smooth. The constant regression coefficients for the cloud-cover variable are also presented in Table 3.3. The coefficient is significant only after 7 in the morning and positive : cloud-cover tends to increase electricity demand in the winter.

Table 3.4 presents estimated factor loadings and related constant terms for the trend and stochastic regression effects. Factor loadings and trend constants are hyperparameters estimated during the likelihood maximization step while constant terms for regression effects are estimated with the Kalman Filter, conditional on the estimated hyperparameters. Estimated factor loadings for the trend are close to 1 for all hours in S^{NF}. Constant terms $\hat{c}_{j(r)}^0$ adjust the trend level, and each $\hat{c}_{j(r)}^k, r = 1, \ldots, 8, k = 1, \ldots, 10, 13, \ldots, 18, j = 2, 3$, adjust the corresponding regression coefficient level. For the trend component, more pooling by reducing the number of factors can be considered in a future model embedding more hours together. For Fourier coefficients, factor loadings are close to 1 in most cases. Exceptions correspond to morning hours 7, 8 and evening hours 16 to 20. Those hours are the ones mostly affected by daylight differences during the year: the yearly pattern is more pronounced and difference between hours is greater. It also explains large values for constant terms associated with evening hours.

Table 3.3: Parameter Estimates of dynamic factor model for French load II
See also Table 3.2 and Table 3.4. Column 2 shows standard deviation estimates for the cooling effect (independent for each hour): $\hat{\sigma}_{11,s}$, $s = 1, \ldots, S$; Column 3 shows the standard deviation estimates of the irregular term in the observation equation: $\hat{\sigma}_{\varepsilon,s}$, $s = 1, \ldots, S$; Column 4 shows the constant regression coefficient estimates $\hat{\beta}_s^{12}$, $s = 1, \ldots, S$ for the cloud-cover with t-values in brackets, bolded when significant at 5% level.

Hour s	$\hat{\sigma}_{11,s}$	$\hat{\sigma}_{\varepsilon,s}$	$\hat{\beta}_s^{12}$(t-value)
0	2.00	137.2	7.4 (0.9)
1	2.00	67.3	2.0 (0.3)
2	2.00	135.3	-12.3 (1.7)
3	2.00	164.6	-6.1 (0.8)
4	2.03	2.1	-10.2 (1.5)
5	2.04	177.6	-4.7 (0.7)
6	2.62	196.9	-10.4 (1.5)
7	2.27	2.6	12.1 (1.7)
8	2.75	224.7	**57.3** (7.9)
9	2.20	171.5	**66.6** (7.7)
10	2.03	28.8	**89.9** (10.8)
11	2.04	195.6	**113.2** (13.2)
12	3.10	187.5	**64.6** (9.3)
13	2.66	75.6	**74.9** (10.5)
14	5.31	178.4	**76.2** (10.1)
15	2.45	172.2	**93.6** (11.3)
16	2.34	70.6	**97.9** (12.2)
17	2.61	217.1	**64.1** (7.2)
18	2.00	187.5	**91.3** (9.3)
19	2.00	1.1	**61.5** (7.2)
20	2.00	200.1	**58.9** (7.7)
21	2.63	210.2	**42.0** (5.7)
22	2.69	118.7	**22.8** (3.2)
23	1.97	163.1	**25.0** (3.4)

Some estimated factor loadings and constants have large positive values for the daytypes, especially for the weekend and for the early Monday. Generally a large factor loading is associated with a large constant term, so that the overall regression effect remains negative for hours in S^{NF}. This identification problem may be explained by the fact that regressors are dummies in this case. As for the trend, except for the early morning hours 1 and 2, factor loadings for the heating effect are very close to 1. More pooling is therefore possible for this regression effect. For the smoothed-regression effect, there is also an identification problem for hours 1 and 2: factor loadings are very large and positive, constant terms are highly negative. The overall effect is similar for the factor hour 0 and for the non-factor hours 1,2. For extra regression effects specific to the

group of three morning hours (6, 7, 8), we notice an identification problem for a Fourier coefficient. Otherwise, factor loadings are close to 1. Wednesday and Thursday effects seem less important for 8 than for 6 and 7 since factor loadings are very small and the associated constants are insignificant.

Finally, Table 3.5, at the end of section 3.4.6, provides likelihood results for each trivariate model. This output can be compared with the maximized likelihood obtained by estimating the corresponding univariate model for all hours, i.e. benchmark model (3.12). Univariate results are grouped for comparison with the factor model. Results shows that the factor model obtains a higher likelihood for all groups of hours. From this point of view, the factor model seems more satisfactory than the univariate modelling of each hour. The total factor model has 86+72+376=534 free parameters as shown in Tables 3.2, 3.3, and 3.4, of which 24+172=196 coefficients are estimated as state variables. Note the remark under (3.7) about the trend component. The total univariate model has 13*24+18=330 free parameters, of which 24 coefficients are estimated as state variables. The total loglikelihoods differ 37728. Note that regular likelihood ratio tests between the models do not apply as they are not nested and have different coefficients included in the state vector.

Table 3.4: Parameter Estimates of dynamic factor model for French load III

See also Table 3.2 and Table 3.3. Parameter estimates of the model in 3.4.2 Top: factor loading estimates for the trend ($\hat{\lambda}_s^0, s \in S^{NF}$) and stochastic regression effects ($\hat{\lambda}_s^k, s \in S^{NF}$, $k = 1, \ldots, 10, 13, \ldots, 18$) ; Bottom: estimated constants for the trend ($\hat{c}_s^0, s \in S^{NF}$) and stochastic regression effects ($\hat{c}_s^k, s \in S^{NF}$, $k = 1, \ldots, 10, 13, \ldots, 18$) with 5 % significant t-values in bold (for regression coefficients only).

	Component	s=1	s=2	s=4	s=5	s=7	s=8	s=10	s=11	s=13	s=14	s=16	s=17	s=19	s=20	s=22	s=23
$\hat{\lambda}_s^0$	Trend	0.99	0.99	1.01	1.04	1.02	1.01	1.00	1.01	1.02	1.03	1.00	1.00	0.98	0.95	0.99	0.98
$\hat{\lambda}_s^1$	$a_{1,t}$	0.98	0.96	0.98	0.98	1.07	0.80	1.00	0.98	1.07	1.07	0.97	0.89	0.94	0.87	1.06	1.09
$\hat{\lambda}_s^2$	$b_{1,t}$	0.99	0.94	0.90	0.68	1.34	0.83	1.08	1.09	1.13	1.17	1.08	1.34	0.08	0.22	0.93	1.05
$\hat{\lambda}_s^3$	$a_{2,t}$	0.94	1.23	0.90	1.06	1.11	0.96	0.94	0.87	1.06	1.01	14.82	11.26	1.08	0.32	1.37	1.24
$\hat{\lambda}_s^4$	$b_{2,t}$	1.00	0.95	0.92	0.74	-2.51	-1.69	1.02	1.01	1.08	1.09	0.99	0.71	1.01	0.93	0.96	0.86
$\hat{\lambda}_s^5$	Monday	3.24	5.04	0.23	-2.35	1.12	0.67	1.09	1.04	0.86	0.75	2.24	2.19	0.74	0.49	0.78	0.15
$\hat{\lambda}_s^6$	Friday	0.82	3.20	-0.04	-1.59	0.73	-0.12	0.83	-0.08	12.10	15.99	1.09	1.04	1.35	1.30	0.94	0.92
$\hat{\lambda}_s^7$	Saturday	4.94	9.76	1.37	2.46	1.16	0.71	0.72	0.54	1.35	1.51	1.11	1.24	0.83	0.67	2.03	1.68
$\hat{\lambda}_s^8$	Sunday	0.40	4.16	1.79	3.95	0.88	0.75	-0.05	-0.37	1.61	1.75	1.04	0.95	7.25	8.35	0.67	0.32
$\hat{\lambda}_s^9$	Heating	0.65	0.64	0.96	0.97	1.02	1.05	0.98	0.96	1.00	0.97	0.99	1.17	1.08	1.18	0.98	0.96
$\hat{\lambda}_s^{10}$	S-Heating	17.27	16.11	0.98	0.91	1.22	1.67	1.31	1.47	1.06	1.11	0.99	0.76	0.89	0.80	0.68	-0.04
$\hat{\lambda}_s^{13}$	$a_{1,t}^{WE}$					6.43	5.23										
$\hat{\lambda}_s^{14}$	$b_{1,t}^{WE}$					1.04	1.12										
$\hat{\lambda}_s^{15}$	$a_{2,t}^{WE}$					1.04	0.85										
$\hat{\lambda}_s^{16}$	$b_{2,t}^{WE}$					0.99	1.01										
$\hat{\lambda}_s^{17}$	Wednesday					1.71	-0.16										
$\hat{\lambda}_s^{18}$	Thursday					1.06	0.08										
$\hat{c}_s^0$	Trend	81	-2021	-386	1289	3533	5577.6	219	430	-1216	-2846	-779	17	-487	-1734	2010	500
$\hat{c}_s^1$	$a_{1,t}$	25	**91**	2	**514**	105	**-436**	**-417**	**-329**	**-77**	**-291**	**644**	**3264**	**-126**	**-1459**	**-428**	**-880**
$\hat{c}_s^2$	$b_{1,t}$	**66**	**142**	**-51**	-2	87	**359**	**-124**	-114	3	**-206**	**-352**	**-347**	**846**	**633**	**297**	**160**
$\hat{c}_s^3$	$a_{2,t}$	**-129**	112	-69	**-752**	**-182**	**-405**	32	149	31	**133**	**6319**	**5741**	**-280**	**-1465**	**861**	**644**
$\hat{c}_s^4$	$b_{2,t}$	24	28	**-86**	**-325**	-354	-80	**213**	**342**	-13	-20	**-197**	**-776**	91	263	**-127**	**-208**
$\hat{c}_s^5$	Monday	**8846**	**16083**	**-2120**	**-10372**	**381**	-15	**221**	**188**	-11	**-233**	**904**	**916**	**53**	**-84**	-16	**-336**
$\hat{c}_s^6$	Friday	11	-132	49	85	-6	52	**-158**	**-553**	**2154**	**2798**	**-95**	**178**	**504**	**676**	**152**	**207**
$\hat{c}_s^7$	Saturday	**1721**	**3897**	**-178**	-137	**-838**	**-3711**	**-2390**	**-3985**	**644**	**1304**	**1279**	**3414**	**-349**	**-829**	**6250**	**5106**
$\hat{c}_s^8$	Sunday	**-2102**	**8711**	**2249**	**8734**	**-5070**	**-6081**	**-13667**	**-17761**	**3994**	**4632**	**933**	**916**	**75447**	**90317**	**-1469**	**-3099**
$\hat{c}_s^9$	Heating	**52**	**49**	**19**	**26**	**-22**	**-47**	**-8**	**-21**	**-13**	**-12**	**10**	**-55**	**-43**	**-97**	**44**	**44**
$\hat{c}_s^{10}$	S-Heating	**-11917**	**-11021**	**-15**	**24**	**-217**	**-595**	**-230**	**-359**	**-39**	**-118**	**11**	**179**	**92**	**186**	**225**	**773**
$\hat{c}_s^{13}$	$a_{1,t}^{WE}$					**4510**	**3766**										
$\hat{c}_s^{14}$	$b_{1,t}^{WE}$					**-224**	16										
$\hat{c}_s^{15}$	$a_{2,t}^{WE}$					**471**	140										
$\hat{c}_s^{16}$	$b_{2,t}^{WE}$					-167	-179										
$\hat{c}_s^{17}$	Wednesday					**-78**	7										
$\hat{c}_s^{18}$	Thursday					**-76**	77										

3.4.4 Signal extraction

Using estimated parameters described in the previous section, we focus on in-sample estimates of the different components of electricity loads: trend and stochastic regression effects. Signal extraction is obtained by the Kalman smoothing algorithm which also computes standard-errors for each component. We first describe the long-term evolution of each effect for the hours in S^F. Then we present in more details signal extraction with associated standard-errors for the morning hour 9. We also compare these results with estimates from the univariate benchmark models in (3.12). Estimates for the first year of the estimation sample are not shown in graphical outputs as they are imprecisely estimated and less relevant.

Daily pattern for all components of electricity load

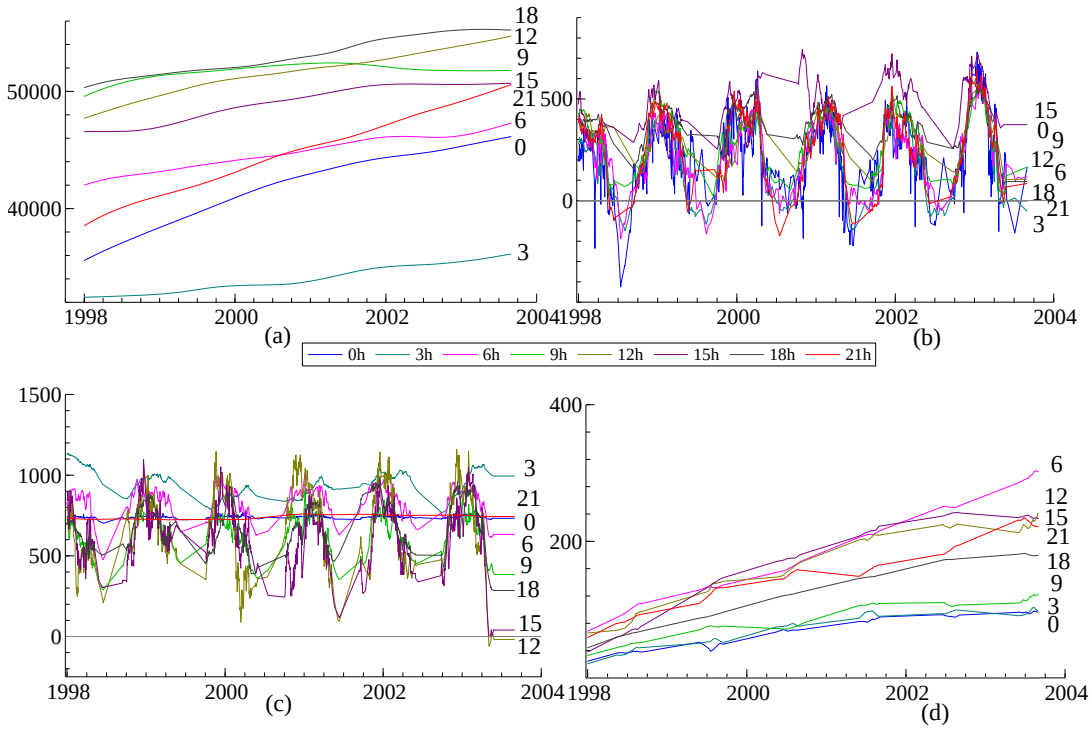

Figure 3.5: French hourly electricity loads - smoothed estimates of: (a) Trend $\hat{\mu}_{s,t}, s = 0, 3, 6, 9, 12, 18, 21$; (b) Heating degrees regression coefficient $\hat{\beta}^9_{s,t}, s = 0, 3, \ldots, 21$; (c) Smoothed-Heating degrees regression coefficient $\hat{\beta}^{10}_{s,t}, s = 0, 3, \ldots, 21$; (d) Smoothed-Cooling degrees regression coefficient $\hat{\beta}^{11}_{s,t}, s = 0, 3, \ldots, 21$. Sample: Jan 1997 - Aug 2003, Graph: Jan 1998 - Aug 2003.

Figure 3.5 shows the time variation in the estimated trend and in the regression coefficients of the main weather effects. Figure 3.5 (a) presents the estimated trend $\hat{\mu}_{s,t}$ for hours $0, 3, \ldots, 21$. The trend is smooth due to the small values imposed for the

variances for the level and slope. The trends are clearly positive, more pronounced for the night hours. The picture indicates that we could reduce the number of dynamic factors and pool more hours for each trend factor. Figure 3.5-(b) presents the estimated heating degrees stochastic regression coefficient $\hat{\beta}^9_{s,t}$ for hours $0, 3, \ldots, 21$. The estimated signal has a clear intrayearly pattern. During the summer, the explanatory variable is zero and the signal is therefore unidentified. Erratic values are obtained during the summer when some cold temperatures occur during the night. Except for the afternoon hours 15 to 17 (not in the graph), the coefficients have a similar pattern for all hours, probably a further dimension reduction of the space of dynamic factors could be considered.

Figure 3.5-(c) shows comparable results for the smoothed heating degrees stochastic regression coefficient $\hat{\beta}^{10}_{s,t}$ for all hours $0, 3, \ldots, 21$. However, for hours 21 to 23 the variance is comparatively small. This coefficient is less affected by occasional cold temperatures in the summer as the regressor contains exponentially smoothed temperatures. Figure 3.5-(d) presents the smoothed cooling degrees coefficient estimates $\hat{\beta}^{11}_{s,t}$ for hours $0, 3, \ldots, 21$. This component is estimated independently for all hours from independent random-walks with hour-specific standard deviations as presented in Table 3.3. Nevertheless, all estimated signals follow a similar positive trend. These cooling effect changes are harder to estimate than heating effect changes since non-zero values for the cooling degrees variable only occur during the summer and only for a few days.

The estimated weekly pattern is given by smoothed estimates of the day-type effects for hours $0, 3, \ldots, 21$ presented in Figure 3.6. The Wednesday and Thursday effects are not shown, as they are only estimated for three hours. Figure 3.6-(a) shows the estimated effect for the Monday day-type $\hat{\beta}^5_{s,t}$ for hours $0, 3, \ldots, 21$. This effect equals the regression coefficient since the Monday variable is a simple dummy variable. The effect is negative for all hours. The larger negative values are obtained for the early morning hours corresponding to the end of the week-end. Estimated values get closer to 0 from midnight to hour 23. The overall effect is almost constant except for the morning hours 9 to 11 and for the evening hour 18 to 20 with a repetitive decrease in August. Figure 3.6-(b) displays the Friday effect, $\hat{\beta}^6_{s,t}$ for for hours $0, 3, \ldots, 21$. The Friday effect clearly differs from the Monday effect. It is not significant for the early morning hours, but during the day it becomes significantly negative as the weekend draws near. The larger negative values correspond to the end of the working day hours 15 to 20. For the late night hours, the effect is smaller again. Overall, the Friday effect does not vary a lot from week to week, but it exhibits a strong intra-yearly pattern for the afternoon hours 15 to 17. Economic activity slows down in August so that the effect of the upcoming weekend is smaller than during the rest of the year.

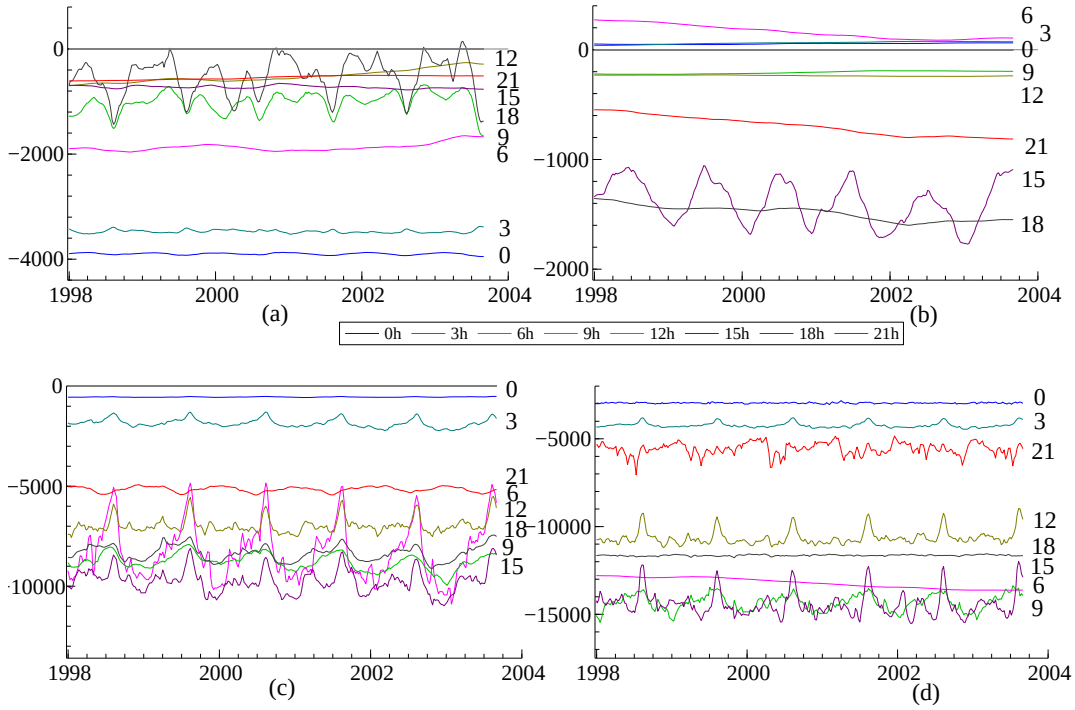

Figure 3.6: French hourly electricity loads - smoothed estimates of: (a) Monday dummy regression coefficient $\hat{\beta}^5_{s,t}, s = 0, 3, \ldots, 21$; (b) Friday dummy regression coefficient $\hat{\beta}^6_{s,t}, s = 0, 3 \ldots, 21$; (c) Saturday dummy regression coefficient $\hat{\beta}^7_{s,t}, s = 0, 3 \ldots, 21$; (d) Sunday dummy regression coefficient $\hat{\beta}^8_{s,t}, s = 0, 3 \ldots, 21$; . Sample: Jan 1997 - Aug 2003, Graph: Jan 1998 - Aug 2003.

Figure 3.6-(c) exhibits the Saturday effect $\hat{\beta}^7_{s,t}$ for hours $0, 3, \ldots, 21$. The effect is highly negative, especially for usual working hours. The signal shows an intra-yearly pattern, more pronounced for working hours with a positive peak in August, confirming that the weekend effect is less important in the summer. Figure 3.6-(d) displays the Sunday effect $\hat{\beta}^8_{s,t}$ for hours $0, 3, \ldots, 21$. The effect is larger than for Saturdays. Morning and evening hours seem to follow opposite dynamics for the Saturday and the Sunday effect.

We do not show the time-varying regression coefficients for the Fourier terms. The Fourier coefficients are highly stochastic and difficult to interpret by themselves. Some coefficients exhibits a strong intra-yearly pattern. Results can be put in perspective with Figure 3.7, which presents the changing yearly components, i.e. the sum consisting of the dynamic trend plus the regression effects due to the Fourier components, $\hat{\mu}_{s,t} + \sum_{k=1}^{4} \left(\hat{\beta}^k_{s,t} x^k_{s,t} \right)$ for each hour s,$s = 0, \ldots, 23$, where extra Fourier components for hours 6, 7 and 8 in the weekends are added. Each panel of Figure 3.7 is associated with one trivariate factor model. There is an upward trend for all hours. During the day, load is minimal around 3-4 in the morning and maximal at 18-19 in the evening. The yearly

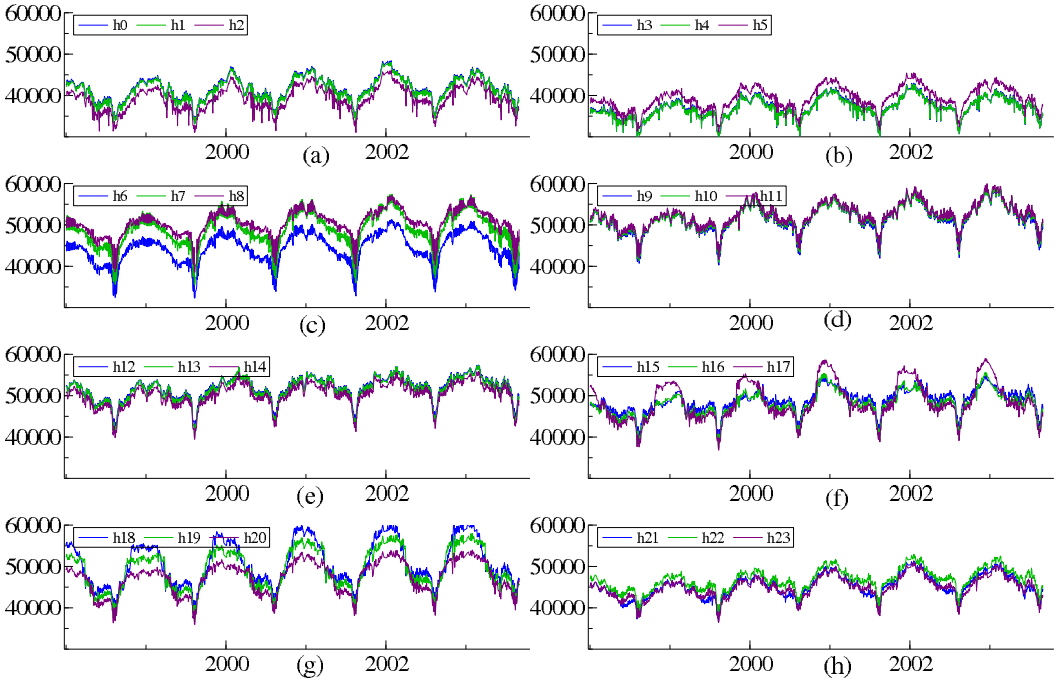

Figure 3.7: French hourly electricity loads - In-sample smoothed estimates of the sum of the trend and the yearly pattern, $\hat{\mu}_{s,t} + \sum_{k=1}^{4} \left(\hat{\beta}_{s,t}^{k} x_{s,t}^{k}\right)$, for hours in Panel (a) $s = 0, 1, 2$; (b) $s = 3, 4, 5$; (c) $s = 6, 7, 8$ (with extra yearly component $\sum_{k=13}^{16} \left(\hat{\beta}_{s,t}^{k} x_{s,t}^{k}\right)$); (d) $s = 9, 10, 11$; (e) $s = 12, 13, 14$; (f) $s = 15, 16, 17$; (g) $s = 18, 19, 20$; (h) $s = 21, 22, 23$. Sample: Jan 1997 - Aug 2003, Graph: Jan 1998 - Aug 2003.

pattern is most prominent for peak hours, in the early morning and in the evening. Time-varying regression coefficients capture the strong decrease of electricity demand in August for all hours of the day. The regular winter increase associated with dark afternoons shows most clearly from hour 17 to 19. The winter increase due to low temperatures is modelled by the heating coefficients, which were presented in Figure 3.5.

Detailed results for 9h and comparison with full independent model

Figure 3.8 presents the smoothed estimates of most stochastic regression coefficients for the multivariate factor model (blue, continuous line) at 9 as well as the stochastic regression coefficients for the corresponding univariate model (red, dashed line). Panel 3.8(h) shows the sum of the trend and yearly Fourier components instead of the coefficients for these components as the separate coefficients are not easy to interpret. Figure 3.9 displays the standard errors corresponding to the estimates of Figure 3.8. At the end of the estimation sample these standard errors are comparatively large as only past observations are available for estimation.

Panels 3.8(a) and 3.8(b) show the regression coefficients corresponding to heating and smoothed-heating degrees variables. Both exhibit a repeating intra-yearly pattern, the

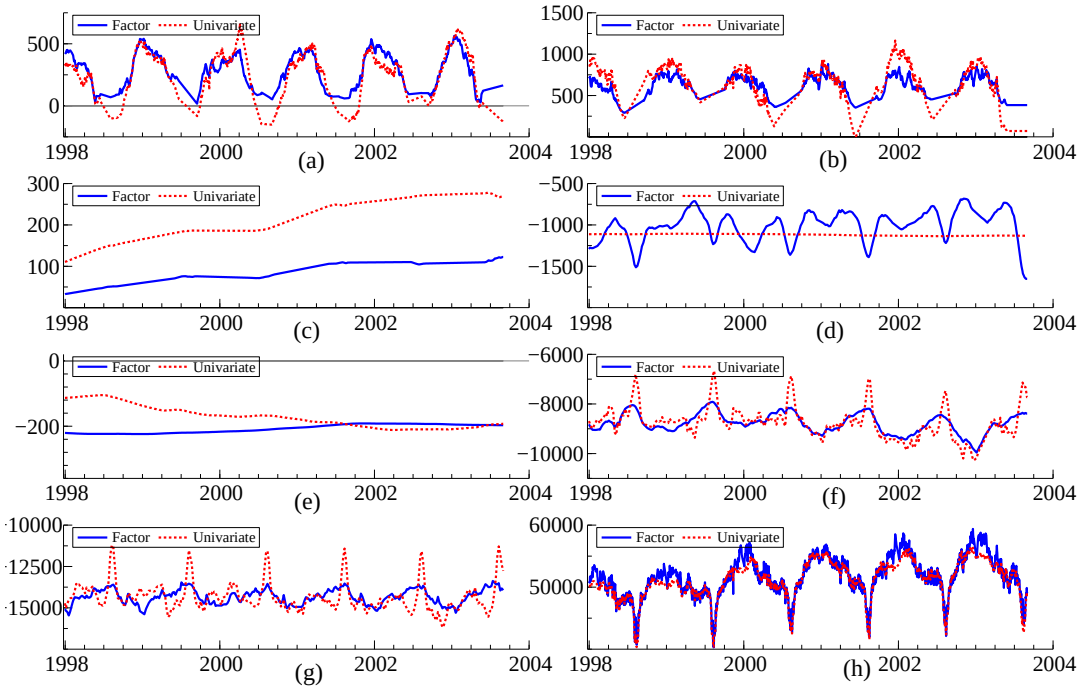

Figure 3.8: French hourly electricity loads - In-sample smoothed estimates of coefficients for hour 9 with multivariate factor (blue, continuous line) and univariate (red, dotted line) models: (a) heating regression coefficient $\hat{\beta}^{9}_{9,t}$; (b) smoothed-heating regression coefficient $\hat{\beta}^{10}_{9,t}$; (c) cooling regression coefficient $\hat{\beta}^{11}_{9,t}$; (d) Monday regression coefficient $\hat{\beta}^{5}_{9,t}$; (e) Friday regression coefficient $\hat{\beta}^{6}_{9,t}$; (f) Saturday regression coefficient $\hat{\beta}^{7}_{9,t}$; (g) Sunday regression coefficient $\hat{\beta}^{8}_{9,t}$; (h) Smoothed estimate of trend plus overall yearly pattern $\hat{\mu}_{9,t} + \sum_{k=1}^{4} \left(\hat{\beta}^{k}_{9,t} x^{k}_{9,t} \right)$; Sample: Jan 1997 - Aug 2003, Graph: Jan 1998 - Aug 2003.

heating coefficient increasing from the start of the heating period to reach its top in the heart of winter and then decreasing until the end of the heating period. Note again that these coefficients estimates are hardly identified in the summer when France experiences long periods with morning temperatures above 15 Celsius degrees. Univariate point estimates of the heating effect - which do not use the information of hours 10 and 11 - are close to the factor model coefficients, but standard errors for the univariate model shown in Panels 3.8(a) and 3.8(b) are clearly larger.

Panel 3.8(c) shows smoothed estimates for the smoothed-cooling degrees coefficient. Both estimates display a steady increase with a consistently larger value for the univariate model. Standard errors of these estimates (especially for the univariate model) are large because most cooling days occur in August, a period when economic activity slows down because of the holidays. It is difficult to distinguish trends in the positive cooling effect from trends in the negative effect due to the holidays. This identification problem requires further investigation.

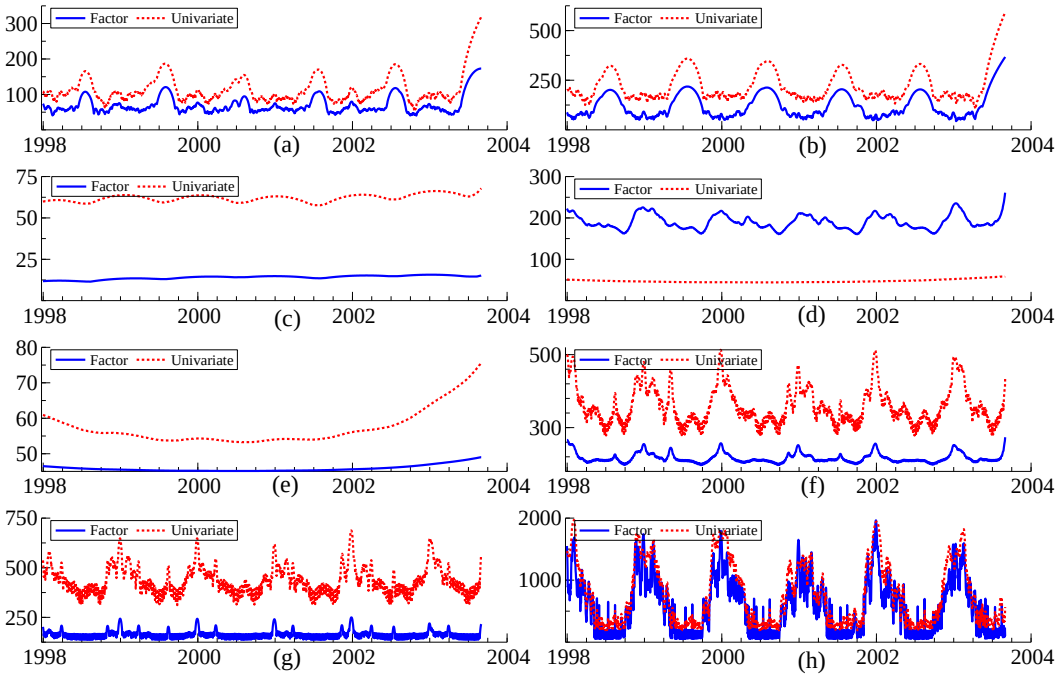

Figure 3.9: French hourly electricity loads - In-sample signal extraction standard errors for hour 9 with multivariate factor (blue, continuous line) and univariate (red, dotted line) models for components: (a) heating regression coefficient $\hat{\beta}^{9}_{9,t}$; (b) smoothed-heating regression coefficient $\hat{\beta}^{10}_{9,t}$; (c) cooling regression coefficient $\hat{\beta}^{11}_{9,t}$; (d) Monday regression coefficient $\hat{\beta}^{5}_{9,t}$; (e) Friday regression coefficient $\hat{\beta}^{6}_{9,t}$; (f) Saturday regression coefficient $\hat{\beta}^{7}_{9,t}$; (g) Sunday regression coefficient $\hat{\beta}^{8}_{9,t}$; (h) Smoothed estimate of trend plus overall yearly pattern $\hat{\mu}_{9,t} + \sum_{k=1}^{4}\left(\hat{\beta}^{k}_{9,t}x^{k}_{k,t}\right)$ Sample: Jan 1997 - Aug 2003, Graph: Jan 1998 - Aug 2003.

Panels 3.8(d) and 3.8(e) show the day-type effect for both models where Tuesday to Friday are considered as the base days of the week. Panel 3.8(d) displays the Monday effect, which is negative for both models. The factor model detects yearly varying values around a mean of -1000 MW. The univariate model estimates a constant value. The difference between the two models does not seem to be significant, taking into account the standard errors presented in Panel 3.9(d). This is the only coefficient with larger standard errors for the factor model. Panel 3.8(e) presents the estimates for the Friday effect. Here the coefficient of the univariate model changes slowly and tends to the nearly constant value -200 MW for the dynamic factor model.

Panel 3.8(f) shows the Saturday effect and (g) the Sunday effect: for both models, stochastic regression coefficients exhibit an intra-yearly pattern, but the factor model estimates are much smoother and miss the peak in August displayed by the univariate model. During the summer, due to less economic activity, there is a smaller difference between electricity consumption on weekdays and weekends. Overall the data show a

trend in the Saturday effect for the winter months from -9000 MW to -10000 MW.

Finally, Panel 3.8(h) shows the overall trend and yearly pattern, as in Figure 3.7. Overall, the results for the univariate and the factor model are similar, but there are interesting differences in the midst of winter when temperatures at 9 are clearly below 15 degrees Celsius. Then the smoothed-heating coefficient in Panel 3.8(b) comes into play and this coefficient is overall larger for the univariate model than for the multivariate factor model. Because the factor model heating effect is smaller, the general seasonal pattern is larger and smoother in the winter. Note from Panel 3.9(h) that the standard errors have the same pattern, with overall larger values for the univariate model. For hour 9, the factor model seems more robust giving narrower confidence intervals.

From all these figures we can see that pooling dynamic effects between neighbouring hours make sense in a serious application. More pooling can be considered: the number of factors for each effect could be different and groups of the number of hours related to each factor as well could also be changed. The one-factor approach for three hours for each component already gives satisfactory results.

3.4.5 In-sample one-step-ahead forecast error analysis

We show the one-step ahead standardised residuals to assess the dynamic specification of our model. For an optimal performance of our model and signal extraction these residuals should not exhibit a clear dynamic structure and their distribution should be approximately Gaussian.

Figures 3.10 and 3.11 show the one-step ahead standardised residuals empirical ACFs respectively for the multivariate model with dynamic factor and for the univariate benchmark models. The dynamic structure is very similar for both models except at hour 1 and 6 where the residuals of the univariate model exhibit a yearly pattern which is not the case for the factor model. The dynamic structure of the residuals is satisfactory for both models although there are some lags for which the autocorrelation is significantly different from zero. The absolute values of the autocorrelations are smaller than 0.2. Both models are satisfactory in this respect.

Figure 3.12 shows the one-step ahead standardised residuals empirical distribution histogram and density estimate (black, continuous line), as well as the corresponding Gaussian distribution (same mean and variance as the empirical density, blue dashed line). We do not show this output for the univariate benchmark model in order to save space. For both models, in the morning hours, the empirical distribution is more concentrated around the mean. Otherwise, the standard Gaussian distribution fits the

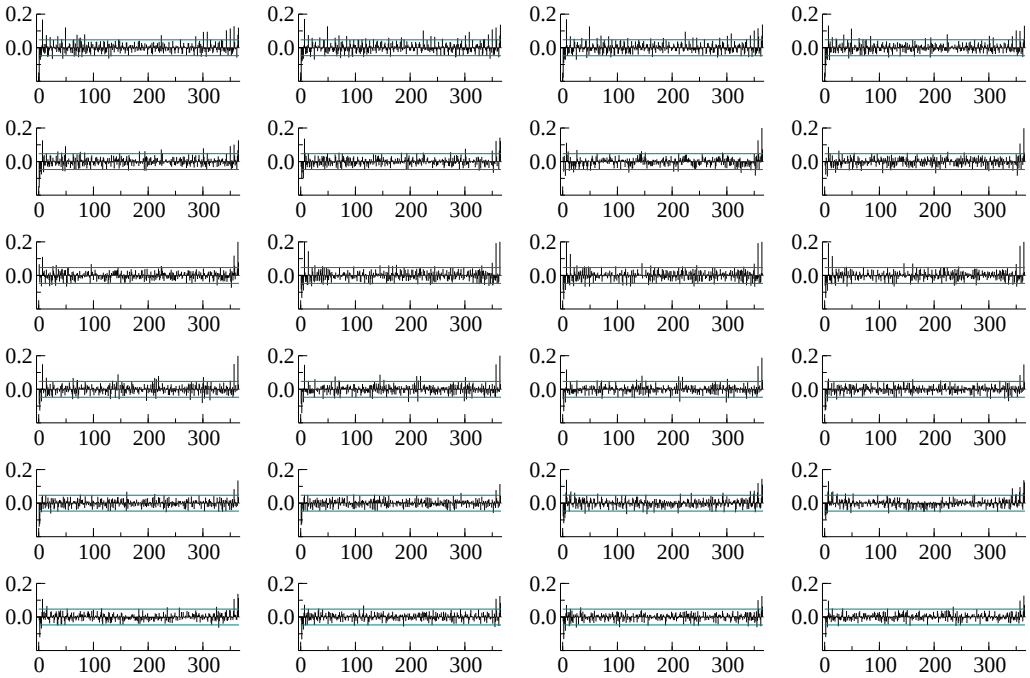

Figure 3.10: French hourly electricity loads - Sample ACF at daily lags of the in-sample standardised residuals from the Dynamic factor model described in 3.4.2, hours 0 to 23, row by row.

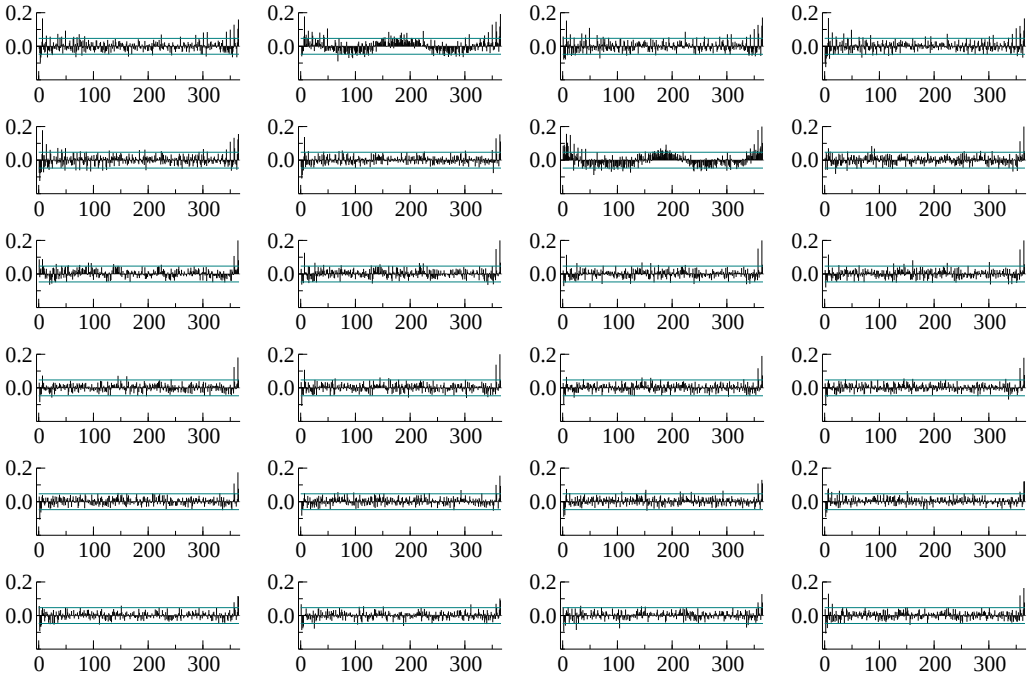

Figure 3.11: French hourly electricity loads - Sample ACF at daily lags of the in-sample standardised residuals ACF from the univariate benchmark models (3.12), hours 0 to 23, row by row.

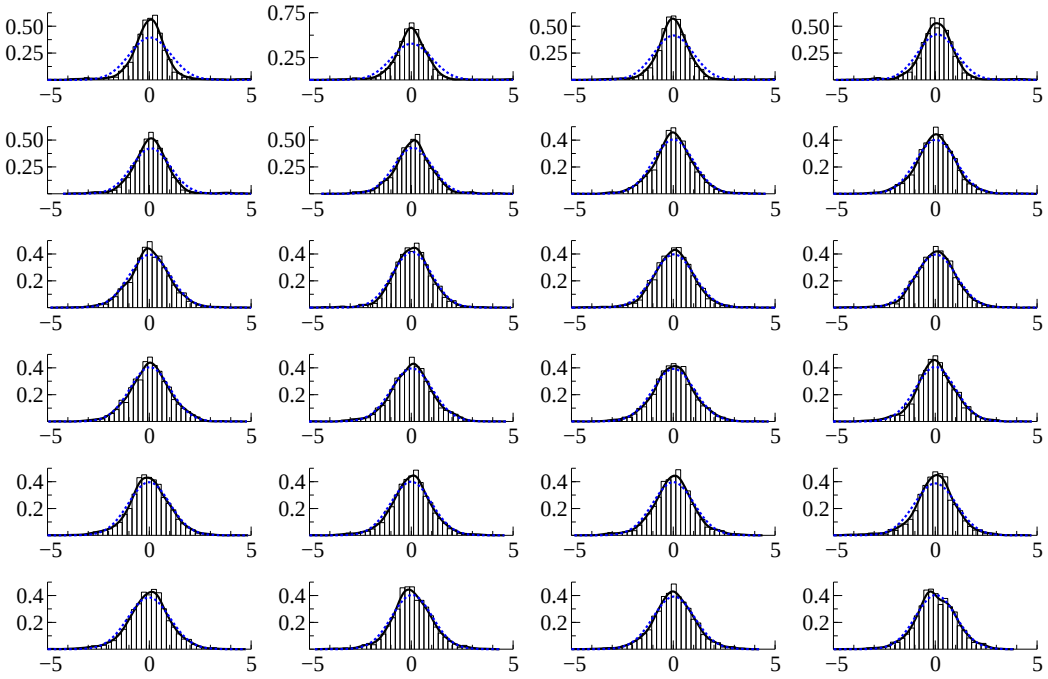

Figure 3.12: French hourly electricity loads - In-sample standardised residuals distribution estimation (histogram, density - black continuous line) and Gaussian approximation (blue, dashed line) for model in 3.4.2, hours 0 to 23, row by row.

residuals empirical distributions reasonably well.

Table 3.5 also shows some hourly diagnostics for the last in-sample year: the residuals sum of squares (RSS) and the standardised residuals sum of squares (SRSS). Regarding the RSS, the Factor model is overall better from 5h to 23h and regarding the SRSS, the Factor model is overall better from midnight to 9 and from 18 to 23 in the evening.

Overall, the dynamic factor model is slightly more satisfactory than our benchmark regarding standardised residuals properties.

3.4.6 Post-sample forecasting results

To evaluate the short-run forecasting accuracy of our dynamic factor model, we compute one-day-ahead hourly forecasts for the prediction period September 1^{st}, 2003 until August 31^{st}, 2004. We run the Kalman filter using the maximum likelihood estimates and the observed values of the explanatory variables on the whole post-sample period to get these forecasts. In Dordonnat et al. (2008), see chapter 2, the effect of using temperature forecasts instead of realized temperatures is studied for related models and similar data, but this did not qualitatively change the conclusions regarding the relative performance of these models. Note that many forecasts are multi-step because of the missing data.

Remark : It is easy to obtain k-day ahead hourly forecasts by running the Kalman filter with the $(k-1)$ previous days treated as missing.

Table 3.5 presents hourly forecasting accuracies for both benchmark univariate model (on the left under UNIV) and dynamic factor model (on the right under FACTOR). The left column under N indicates the number of days actually forecast. Missing days correspond to EJP days, December 23^{rd} until January 3^{rd}, bank holidays, bridge days, and daylight savings days (last Sunday of October and last Sunday of March). In this table, we present two usual measures of accuracy: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The RMSE is given in MegaWatts (MW) and the MAPE in %.

The results are overall very similar for both models. The best forecasting accuracy is obtained for the night hours 21 to 23 for both models. The worst forecasting accuracies are obtained for the hours 7, and 15 to 18, again for both models.

The multivariate factor model is slightly better than the univariate model for the morning hours. We also compared the forecasting performance across days of the week and across months of the year. For hour 9 the MAPE of the factor model varies from 0.93 (0.99) in October to 2.16 (1.94) in August and from 1.08 (1.04) for normal weekdays to 1.77 (1.76) for Saturdays (univariate MAPEs in parentheses). For hour 19, the MAPE varies from 0.87 (0.83) in February to 2.17 (2.04) in April and from 1.23 (1.15) on normal weekdays to 1.84 (1.73) on Mondays. The (differences in) forecasting accuracy strongly depend on the type of hour and day one is forecasting.

It is interesting to notice that using a dynamic factor structure to pool similar hours does not depreciate forecasting accuracy, although it is more constrained than the univariate modelling of each hour.

Table 3.5: Likelihoods and hourly diagnostics for French loads modelling and forecasting

Likelihoods for estimation sample (Lik.), Residuals Sum of Squares (RSS), Standardised Residuals Sum of Squares (SRSS) and post-sample Root Mean Squared Error (RMSE) and MAPE (Mean Absolute Percentage Error) for one-day ahead forecasts for benchmark model (3.12) (left - UNIV) and dynamic factor model of section 3.4.2 (right - FACTOR), with parameter vectors $c^k, k = 1, \ldots, 18$ in the state vector, see also appendix 3.6. N is the number of days actually forecast in post-sample.

		UNIV					FACTOR				
Hour	N	Lik.	RSS	SRSS	RMSE	MAPE	Lik.	RSS	SRSS	RMSE	MAPE
0	319		3.51E+08	297	958	1.43		3.83E+08	310	986	1.40
1	319		2.95E+08	312	951	1.38		3.03E+08	294	1012	1.44
2	319	-51256	2.76E+08	313	919	1.46	-45735	2.87E+08	282	991	1.50
3	319		2.44E+08	308	910	1.46		2.33E+08	280	906	1.42
4	319		2.01E+08	311	856	1.40		1.92E+08	277	862	1.40
5	319	-50043	1.67E+08	294	792	1.34	-44684	1.64E+08	266	812	1.35
6	319		2.90E+08	373	970	1.52		2.30E+08	318	948	1.46
7	319		3.24E+08	346	1059	1.45		2.94E+08	335	1026	1.42
8	319	-50217	2.81E+08	310	958	1.29	-46436	2.48E+08	305	949	1.30
9	319		2.61E+08	314	886	1.21		2.66E+08	300	842	1.19
10	319		2.76E+08	300	892	1.18		2.75E+08	317	884	1.22
11	319	-50705	2.75E+08	297	878	1.16	-45274	2.86E+08	324	912	1.25
12	319		2.45E+08	315	879	1.17		2.40E+08	307	903	1.24
13	319		2.84E+08	309	951	1.30		2.90E+08	321	974	1.38
14	319	-50666	2.95E+08	287	982	1.38	-45203	3.02E+08	303	1014	1.48
15	319		2.92E+08	272	1041	1.47		2.96E+08	262	1092	1.52
16	319		3.03E+08	295	1041	1.52		3.12E+08	290	1123	1.61
17	319	-51264	2.96E+08	303	1038	1.50	-46239	2.99E+08	284	1122	1.58
18	319		3.04E+08	329	1020	1.46		3.26E+08	325	1067	1.55
19	319		3.00E+08	362	978	1.34		3.19E+08	363	1040	1.47
20	319	-50602	2.24E+08	351	823	1.22	-47235	2.31E+08	355	849	1.30
21	319		1.64E+08	355	730	1.13		1.53E+08	343	728	1.09
22	319		1.49E+08	342	693	1.02		1.46E+08	352	708	1.03
23	319	-48797	1.37E+08	325	654	1.02	-45016	1.32E+08	320	667	1.04
Total		-403550					-365822				

3.5 Conclusion

We use the multivariate linear Gaussian state-space framework to model time series of periodic high frequency data. We consider the analysis of different dynamic factors for different components, not only in trends and cycles, but also in stochastic regression coefficients. The dynamic factors provide a parsimonious specification of the different common dynamics in the different components. The investigation of changing weather effects and weekly patterns for hourly electricity loads motivates our analysis.

A small scale Monte Carlo experiment with a regressor measuring the heating effect in electricity load modelling is performed. The results illustrate the theoretical advantage in terms of signal extraction accuracy of the dynamic factor specification for time-varying regression coefficients.

We develop our model and methods in an empirical study of French hourly electricity loads (1997-2003). We test our model out-of-sample for 2003-2004. The empirical model for daily time series of vectors of hourly loads specifies dynamic components for trends, yearly patterns, day types and temperature effects, with independent dynamic factors for consecutive groups of hours in the day. Some dynamic components exhibit a smooth variation over the long period under study while others display an intra-yearly pattern slowly evolving through years. Satisfactory residual-based diagnostics are obtained for the loads of all 24 hours of the day. We evaluate the forecasting performance of the factor model and compare it with univariate models. The factor model gives satisfactory results for one-day-ahead out-of-sample forecasting.

3.6 Appendix: State-space form of the periodic dynamic regression model with factors

We show how model (3.1)-(3.2)-(3.3)-(3.4) with a single dynamic factor for each dynamic component can be put in state-space form.

We first define the transition equation for the common trend. Then we define the transition equation for the factor in each time-varying regression component and finally we present the measurement equation.

The state vector is: $\alpha_t = \begin{pmatrix} \alpha_t^{0\prime} & \alpha_t^{1\prime} & \dots & \alpha_t^{K\prime} \end{pmatrix}'$.

3.6.1 Transition equation for a single common trend

The vector of fixed coefficients $(c_2^0 \dots c_S^0)'$ can be estimated recursively by including it in the state vector. Let

$$\alpha_t^0 = \begin{pmatrix} c_2^0 \dots c_S^0 & g_t^0 & f_t^0 \end{pmatrix}'$$

The transition equation may be written as follows:

$$\alpha_{t+1}^0 = \begin{pmatrix} I_{S-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \alpha_t^0 + \begin{pmatrix} 0_{S-1} \\ w_t^0 \\ v_t^0 \end{pmatrix}$$

If c^0 is estimated as a hyperparameter in the likelihood maximization step then the specification becomes:

$$\alpha_t^0 = \begin{pmatrix} g_t^0 & f_t^0 \end{pmatrix}', \quad \alpha_{t+1}^0 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \alpha_t^0 + \begin{pmatrix} w_t^0 \\ v_t^0 \end{pmatrix}$$

We use the latter approach in the empirical application, where $S = 3$.

3.6.2 Transition equation for a dynamic factor in time-varying regression framework

For each regression effect $k = 1, \dots, K$, the vector of fixed coefficients $\begin{pmatrix} c_2^k & \dots & c_S^k \end{pmatrix}$ is estimated recursively by including it in the state vector. Define

$$\alpha_t^k = \begin{pmatrix} c_2^k & \dots & c_S^k & f_t^k \end{pmatrix}'$$

The transition equation of (3.5) is then written as follows:

$$\alpha_{t+1}^k = \begin{pmatrix} I_{S-1} & 0 \\ 0 & 1 \end{pmatrix} \alpha_t^k + \begin{pmatrix} 0_{S-1} \\ e_t^k \end{pmatrix} \tag{3.13}$$

We use this approach in the empirical application with $S = 3$.

3.6.3 Measurement equation for the general model

When the vector c^0 is included in the state vector as in (3.13) above, the measurement equation of (3.5) for the general model becomes:

$$y_t = \begin{pmatrix} 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & x_{1,t}^1 & & 0 & \dots & 0 & x_{1,t}^K \\ 1 & & & \vdots & \lambda_2^0 & x_{2,t}^1 & & & \lambda_2^1 x_{2,t}^1 & & x_{2,t}^K & & & \lambda_2^K x_{2,t}^K \\ & \ddots & & \vdots & \vdots & & \ddots & & \vdots & \dots\dots & & \ddots & & \vdots \\ & & 1 & 0 & \lambda_S^0 & & & x_{S,t}^1 & \lambda_S^1 x_{S,t}^1 & & & & x_{S,t}^K & \lambda_S^K x_{S,t}^K \end{pmatrix} \alpha_t + \varepsilon_t$$

When c^0 is a part of the "hyperparameters" of the model, the measurement equation is written as:

$$y_t = \begin{pmatrix} 0 \\ c_2^0 \\ \vdots \\ c_S^0 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & x_{1,t}^1 & & 0 & \dots & 0 & x_{1,t}^K \\ \vdots & \lambda_2^0 & x_{2,t}^1 & & & \lambda_2^1 x_{2,t}^1 & & x_{2,t}^K & & & \lambda_2^K x_{2,t}^K \\ \vdots & \vdots & & \ddots & & \vdots & \dots\dots & & \ddots & & \vdots \\ 0 & \lambda_S^0 & & & x_{S,t}^1 & \lambda_S^1 x_{S,t}^1 & & & & x_{S,t}^K & \lambda_S^K x_{S,t}^K \end{pmatrix} \alpha_t + \varepsilon_t.$$

The latter approach is used in the empirical application.

Chapter 4

Intradaily smoothing splines for time-varying regression models of hourly electricity load

Abstract In this chapter we develop a methodology for the modelling of hourly electricity load with focus on the intradaily pattern. We use cubic splines for a smooth modelling of the electricity demand intradaily pattern. The spline weight matrix is known a priori by setting the number of knots and their hour-of-the-day position. The other seasonal effects for electricity load (weekly and yearly pattern) as well as weather influence are taken care of by means of a dynamic regression model where regression coefficients follow random-walk processes. The trend is stochastic as well. The model fits in the multivariate linear Gaussian state-space framework. Considering the vast amount of data (long time series and many regressors), general maximum likelihood estimation of the state-space model is intractable. We manage the computation time issue for parameter estimation by following the method of Jungbacker & Koopman (2008) that implies transforming the model equations to reduce data dimension with no loss of information. Hourly explanatory variables are also transformed. We apply the method on a ten-year dataset of French hourly electricity demand. We analyse the evolution of the different dynamic components over the estimation sample and we further study the intradaily pattern of these components. We compare the model results for the intradaily patterns with the corresponding univariate models and the block diagonal factor model of chapter 3. The model gives satisfactory insights for daily explanatory variables effects, consistent with benchmark models as well as with the internal EDF model. We also suggest further empirical improvements of the model for hourly electricity load.

4.1 Introduction

This chapter develops a statistical methodology within the state-space framework to model the daily curve for electricity demand. The load curve itself is modelled using cubic regression splines and dynamic regression models are used to model the dynamics through days of the week and years. Hourly electricity demand is considered as high frequency data, depending on multiple seasonal patterns evolving on a long time scale. We are mainly concerned with the modelling of the daily pattern but its form is evolving both a on weekly and a yearly basis. An effective model requires to take these features into account. Moreover, the influence of weather conditions needs to be considered for electricity demand.

Statistical modelling for hourly (or even higher-frequency) electricity demand is popular. Taylor (2003) adopts univariate exponential smoothing methods for short-term forecasting during periods without weather effect (heating or cooling). A multivariate Bayesian approach is adopted by Smith (2000) and Cottet & Smith (2003). A model for French data developed by the main electricity producer in France is described in Bruhns et al. (2005). A more thorough review is provided by Bunn & Farmer (1985) and by Lotufo & Minussi (1999).

Our modelling approach for hourly electricity loads involves two main features. First we consider the data as a vector of daily loads so that we model a 24×1 time series vector y_t, thereby adopting a periodic method where dynamic properties depend on the hour of the day. Then we model daily loads with smoothing splines where the number of knots and their abscissa (hour-of-the-day) position are fixed a priori. The second feature is the modelling of the load coordinates of the knots with multivariate dynamic regression models. In short, we build a dynamic factor model where factors are spline knots and factor loadings correspond to spline weights. Our work can be put in perspective with the time-varying regression splines model of Harvey & Koopman (1993) who modelled the intraweekly pattern of electricity load in a univariate model. A multivariate stochastic spline approach was adopted by Koopman & Ooms (2003) to model the intramonthly variation of daily data of tax revenues.

The model fits in the multivariate linear Gaussian state-space framework. The state-space methodology approach for hourly electricity demand has successively been adopted in Dordonnat et al. (2008), see also chapter 2, and Dordonnat et al. (2009), see also chapter 3. While the former study builds a periodic dynamic regression model where each hour has its own dynamics, and where dynamics related to a specific component can be correlated without further constraints, the latter study imposes restrictions through

dynamic factors both in structural components and in stochastic regression coefficients, for groups of hours. The block structure of these models implies abrupt changes between submodels. The spline structure we choose for daily loads in this chapter smoothes these changes. Moreover, we model all hours simultaneously instead of groupwise.

The multivariate linear Gaussian state-space framework embeds a wide range of classical statistical time series models: the classical ARIMA, structural models of Harvey (1989), time-varying and constant regression models as well as exponential smoothing methods. The framework deals with univariate as well as multivariate time series, stationary or not. The different classical methodologies can be considered all together as long as identification is preserved. All the components are modelled separately and permit separate interpretation of their dynamics which is of high importance in the case of multiple seasonal patterns. This framework allows an easy treatment of missing data and forecasts are computed straightforwardly. The loglikelihood is computed using the well-known Kalman filtering algorithm via prediction error decomposition and is maximized to estimate unknown parameters. Numerical methods like the EM algorithm can be used for likelihood maximization. However, in the case of long time series of high frequency data computation time can become problematic. To tackle this problem we follow the dimension reduction method for dynamic factor models described in Jungbacker & Koopman (2008) to compute the loglikelihood function and for signal extraction. The spline model is applied to French national hourly electricity demand and compared with two other state-space specifications: univariate dynamic regression models and periodic dynamic regression models with dynamic factors as in Dordonnat et al. (2009), see also chapter 3. A recent macroeconomic application of a related model can be found in Bowsher & Meeks (2008) who build a low dimension VAR model for the yield curve using a cubic spline approach. They also suggest a knot selection strategy.

Our main goal is to investigate the effect of the different components of electricity demand and the common features in the dynamics of the intradaily load curve over the years during a long period. The state-space methodology also allows straightforward computations for post-sample one-day ahead forecasts. In this chapter however, forecasting accuracy is a secondary issue.

The remainder of the chapter is organized as follows: section 4.2 describes the model and parameter estimation methods. Section 4.3 details the dataset and the final model for French national hourly electricity load. Empirical results are discussed in section 4.4, both from an estimation and from a forecasting standpoint. Section 4.5 concludes.

4.2 Methodology

In this chapter we present a multivariate dynamic regression model in the linear Gaussian state-space framework for high frequency data. We introduce smoothing splines to reduce the dimensions of the dynamics in our model and reach a parsimonious specification. We first present the general model specification then we discuss the parameter estimation issue.

4.2.1 General Model

We build a time series regression model for a univariate time series subject to multiple seasonal patterns. We denote by S the shorter seasonal frequency and transform the original univariate time series into a $S \times 1$ time series vector $y_t = (y_{1,t}, \ldots, y_{S,t})'$. Then we write a periodic dynamic regression model for y_t as follows:

$$y_t = \mu + W\lambda_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \Sigma_\varepsilon), \quad t = 1, \ldots, T \tag{4.1}$$

In equation (4.1), the $S \times 1$ vector $\mu = (\mu_1, \ldots, \mu_S)'$ is a constant unknown level term. The S elements of y_t depend on the $R \times 1$ dynamic vector of reduced dimension $\lambda_t = (\lambda_{1,t}, \ldots, \lambda_{R,t})'$. W is the $S \times R$ spline weight matrix so that λ_t is a vector of time-varying knots coefficients. We give details on the spline weights computation in the Appendix. We assume that the number of knots R as well as the knot positions are known a priori so that W is a known constant matrix. Finally, the $S \times 1$ vector $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{S,t})'$ is a zero mean Gaussian disturbance with unknown covariance matrix Σ_ε. We suppose Σ_ε is diagonal: $\Sigma_\varepsilon = \text{diag}\left(\left(\sigma^2_{\varepsilon,s}\right)_{s=1,\ldots,S}\right)$ so that the different time series in y_t are correlated only through their respective dependence on knot vector λ_t.

The time series y_t depends on K exogenous variables through λ_t. Two kinds of variables are considered: K_1 variables have a lower frequency than $y_{s,t}$ so that their values change with t, not with s. These explanatory variables are therefore denoted as scalars $x^k_t, k = 1, \ldots, K_1$. The other $(K - K_1)$ variables change with s and are placed in the $R \times 1$ vectors $X^k_t = \left(x^k_{1,t}, \ldots, x^k_{S,t}\right)'$ for $k = K_1 + 1, \ldots, K$.

The knot coefficient vector λ_t follows a dynamic regression model:

$$\lambda_t = \mu^*_t + \sum_{k=1}^{K_1} x^{*k}_t \beta^{*k}_t + \sum_{k=K_1+1}^{K} X^{*k}_t \beta^{*k}_t, \quad t = 1, \ldots, T \tag{4.2}$$

In equation (4.2), $R \times 1$ vector $\mu^*_t = \left(\mu^*_{1,t}, \ldots, \mu^*_{R,t}\right)'$ is a stochastic trend. λ_t also depends on K explanatory variables. Each $R \times R$ diagonal matrix X^{*k}_t derives from the

$S \times 1$ vector X_t^k. The precise construction is detailed below. The scalar variables x_t^k are not transformed: $x_t^{*k} = x_t^k, k = 1, \ldots, K_1$. The $R \times 1$ vectors $\beta_t^{*k} = \left(\beta_{1,t}^{*k}, \ldots, \beta_{R,t}^{*k}\right)', k = 1, \ldots, K, t = 1, \ldots, T$, contain the corresponding stochastic regression coefficients.

The vector μ_t^* follows a multivariate local linear trend model:

$$\left\{ \begin{array}{rcll} \mu_{t+1}^* & = & \mu_t^* + \nu_t^* + \eta_t^*, & \eta_t^* \sim N(0, \Sigma_\eta) \\ \nu_{t+1}^* & = & \nu_t^* + \zeta_t^*, & \zeta_t^* \sim N(0, \Sigma_\zeta) \end{array} \right. , \quad t=1,\ldots,\mathrm{T}. \tag{4.3}$$

In equation (4.3), the $R \times 1$ vector $\nu_t^* = \left(\nu_{1,t}^*, \ldots, \nu_{R,t}^*\right)'$ is a stochastic slope term, while the $R \times 1$ vectors $\eta_t^* = \left(\eta_{1,t}^*, \ldots, \eta_{R,t}^*\right)'$ and $\zeta_t^* = \left(\zeta_{1,t}^*, \ldots, \zeta_{R,t}^*\right)'$ are zero mean Gaussian disturbances with respective covariance matrices Σ_η and Σ_ζ. Different specifications can be considered for these covariance matrices. Setting them to zero results in deterministic trends for λ_t. Setting only Σ_ζ to zero gives random-walks with constant drift. Setting only Σ_η to zero gives integrated random-walks. If both matrices are non-zero, they can be unrestricted positive semi-definite or different restrictions can be considered for parsimony. We assume in this chapter that Σ_η and Σ_ζ are diagonal matrices: $\Sigma_\eta = \mathrm{diag}\left(\left(\sigma_{\eta,r}^2\right)_{r=1,\ldots,R}\right)$ and $\Sigma_\zeta = \mathrm{diag}\left(\left(\sigma_{\zeta,r}^2\right)_{r=1,\ldots,R}\right)$, so that each knot follows an independent stochastic trend.

The stochastic regression coefficient vectors $\beta_t^{*k}, k = 1, \ldots, K$ follow multivariate random walks:

$$\beta_{t+1}^{*k} = \beta_t^{*k} + e_t^{*k}, \qquad e_t^{*k} \sim N(0, \Sigma_k), \quad t = 1, \ldots, T \tag{4.4}$$

where the $R \times 1$ vectors $e_t^{*k} = \left(e_{1,t}^{*k}, \ldots, e_{R,t}^{*k}\right)', k = 1, \ldots, K$ are zero mean Gaussian noise terms with respective covariance matrices Σ_k. Setting $\Sigma_k = 0$ makes the corresponding regression coefficient constant over time: $\beta_t^{*k} = \beta^{*k}$. Non-zero Σ_k can be unrestricted positive semi-definite or restrictions can be considered as well. We assume that all non-zero matrices Σ_k are diagonal: $\Sigma_k = \mathrm{diag}\left(\left(\sigma_{k,r}^2\right)_{r=1,\ldots,R}\right)$ so that the stochastic regression coefficients for the knots are independent for each $r, r = 1, \ldots, R$.

Equations (4.1) to (4.4) form our smoothing spline dynamic regression model. Note that the specification for the knot vector λ_t can also include structural components such as defined in Harvey (1989) or ARMA components, see Box & Jenkins (1970). Note also that the different irregular terms $\varepsilon_t, \eta_t^*, \zeta_t^*, e_t^{*k}, k = 1, \ldots, K$ are mutually and temporally independent. The initialization of the dynamics $\mu_t^*, \eta_t^*, \beta_t^{*k}, k = 1, \ldots, K$ and constant μ is briefly discussed below.

Before considering model estimation, we summarize the different restrictions in our empirical study:

- dimension R of the knots vector λ_t. The knot positions are selected a priori,

- the covariance matrix Σ_ε in (4.1) is diagonal,
- the covariance matrices $\Sigma_\eta, \Sigma_\zeta$ in the local linear trend equation (4.3) are diagonal,
- the covariance matrices $\Sigma_k, k = 1, \ldots, K$, for the regression coefficients in the random walk equation (4.4) can be zero (constant regression coefficients) or diagonal (independent stochastic regression coefficients).

Remark The model also embeds the general dynamic periodic regression model by taking $R = S$, so that matrix W becomes an identity matrix. In this case, each $y_{s,t}$ depends on its own dynamic factor $\lambda_{s,t}, s = 1, \ldots, S$. Dependence relations between the different time series in y_t depend on the dynamic model specification of vector λ_t. Independent $\lambda_{s,t}, s = 1, \ldots, S$ correspond to the univariate (independent) modelling of each $y_{s,t}$.

4.2.2 Model estimation

We consider maximum likelihood estimation of the unknown parameters in model (4.1)-(4.2)-(4.3)-(4.4):

- the (diagonal) covariance matrix Σ_ε in (4.1),
- the (diagonal) covariance matrices $\Sigma_\eta, \Sigma_\zeta$ in (4.3), and
- the (diagonal) covariance matrices $\Sigma_k, k = 1, \ldots, K$ in (4.4), if not zero.

We first discuss the general estimation method within the multivariate linear Gaussian state-space framework. Then we discuss the application of the dimension reduction method from Jungbacker & Koopman (2008). We also discuss the special estimation of the constant level vector μ in (4.1).

General state-space method

The multivariate linear Gaussian state-space model is written as follows:

$$\left\{ \begin{array}{lcll} y_t & = & Z_t\alpha_t + \varepsilon_t & , \quad \varepsilon_t \sim N(0, H_t), \\ \alpha_{t+1} & = & T_t\alpha_t + R_t\eta_t & , \quad \eta_t \sim N(0, Q_t), \end{array} \right. \quad t=1,\ldots,\mathrm{T} \tag{4.5}$$

The first equation in (4.5) is the measurement equation. The vector of observations y_t depends on the state vector α_t through the measurement matrix Z_t and ε_t is zero mean Gaussian noise with covariance matrix Σ_ε. The second equation in (4.5) is the transition equation for the state vector α_t. T_t is the transition matrix, usually sparse and η_t is zero mean Gaussian noise with covariance matrix Σ_η. R_t is an error loading matrix: if an

element in α_t is deterministic then the corresponding row in R_t is a null vector while the row is a unit vector if the element is stochastically time-varying. The initial state vector distribution is denoted as $\alpha_1 \sim N(a_1, P_1)$. The state-space framework is described in detail in textbooks such as Harvey (1989) or Durbin & Koopman (2001).

Given the model parameters for $Z_t, T_t, R_t, \Sigma_\varepsilon, \Sigma_\eta$, the Kalman filtering algorithm outputs $a_t = \mathrm{E}\left(\alpha_t | y_1, \ldots, y_{t-1}\right)$, the forecast of the state vector α_t based on the past values $y_1, \ldots, y_{t-1}$ and the corresponding variance $P_t = \mathrm{Var}\left(\alpha_t | y_1, \ldots, y_{t-1}\right)$ for $t = 2, \ldots, T$. The forecast error is defined by $v_t = y_t - Z_t a_t$ and the corresponding variance by $F_t = Z_t P_t Z_t' + \Sigma_\varepsilon$. The loglikelihood of the data $l(y) = l\left(y_1, \ldots, y_T\right)$ is deduced from the Kalman filter results using the prediction error decomposition:

$$\log L(y) = -\frac{Tk}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\left(\log|F_t| + v_t' F_t^{-1} v_t\right) \tag{4.6}$$

To estimate unknown parameters, the loglikelihood can be maximized using the EM algorithm, see Shumway & Stoffer (1982) or using a scoring algorithm, where the score vector is evaluated analytically, following Koopman & Shephard (1992), or numerically. An intermediate strategy consists in using the EM algorithm to get close to the optimum values, then switch to a scoring algorithm to avoid the low convergence rate of EM in the late iterations.

Model (4.1)-(4.2)-(4.3)-(4.4) presented in section 4.2.1 can easily be written in state-space form with:

$$\begin{aligned}
&\alpha_t = \begin{pmatrix} \mu & \mu_t^* & \nu_t^* & \beta_t^{*1} & \ldots & \beta_t^{*K} \end{pmatrix}', \\
&Z_t = \begin{pmatrix} I & W & 0 & Wx_t^{*1} & \ldots & Wx_t^{*K_1} & WX_t^{*K_1+1} & \ldots & WX_t^{*K} \end{pmatrix}, \\
&H_t = H = \Sigma_\varepsilon, R_t = \begin{pmatrix} 0 & I_{R(2+K)} \end{pmatrix}', \quad \eta_t = \begin{pmatrix} \eta_t^* & \zeta_t^* & e_t^{*1} & \ldots & e_t^{*K} \end{pmatrix}', \\
&T_t = \begin{pmatrix} I_S & & & & \\ & I_R & I_R & & \\ & & I_R & & \\ & & & I_{R\times K} \end{pmatrix}, \qquad Q_t = Q = \begin{pmatrix} \Sigma_\eta & & & & \\ & \Sigma_\zeta & & & \\ & & \Sigma_1 & & \\ & & & \ddots & \\ & & & & \Sigma_K \end{pmatrix}
\end{aligned} \tag{4.7}$$

Initialization of the state vector α_t: For nonstationary components in the state vector α_t, Durbin & Koopman (2001) discuss the diffuse initialization method. The authors give the corresponding adjustments in the Kalman filtering equations and (diffuse) likelihood evaluation. We use this method for the constant level vector μ, for the (stochastic) regression coefficients in (4.4) and for the slope term ν_1^* in (4.3). The level term in the local linear trend equation (4.3) is initialized to a null constant to identify the constant vector μ.

Estimation of the constant level vector μ and hyperparameters in Q: Within the general state-space method, the vector μ is included in the state vector α_t so that it is estimated recursively using the Kalman filter with the hyperparameters estimated by the likelihood maximization. The resulting diffuse likelihood is also known as the marginal likelihood for the covariance matrices in (4.7), as defined in Jungbacker & Koopman (2008) and references therein.

Dimension reduction

Jungbacker & Koopman (2008) suggest a dimension reduction method for model (4.5) to reduce computation time for the loglikelihood evaluation. We follow their method for model (4.1)-(4.2)-(4.3)-(4.4). The spline weight matrix W in equation (4.1) meets the conditions of Jungbacker & Koopman (2008)-section 3.1.1 so that we can transform the $S \times 1$ vector y_t into the $R \times 1$ vector $y_t^* = \left(y_{1,t}^*, \ldots, y_{R,t}^*\right)'$ using the $S \times R$ transformation matrix P:

$$y_t^* = P' y_t, \qquad P' = \left(W' \Sigma_\varepsilon^{-1} W\right)^{-1} W' \Sigma_\varepsilon^{-1}, \qquad t = 1, \ldots, T. \tag{4.8}$$

We also transform the constant vector μ into $\mu^* = P'\mu$ and each $R \times R$ matrix of transformed explanatory variable X_t^{*k} is deduced from the $S \times 1$ vector X_t^k using $X_t^{*k} = \text{diag}(P' X_t^k)$.

We then model the low-dimensional vector y_t^* instead of y_t:

$$\begin{array}{rcl} y_t^* = & \mu^* + \lambda_t + u_t & \\ = & \mu^* + \mu_t^* + \sum_{k=1}^{K_1} x_t^{*k} \beta_t^{*k} + & \\ & \sum_{k=K_1+1}^{K} X_t^{*k} \beta_t^{*k} + u_t, \qquad u_t \sim NID\left\{0, (W' \Sigma_\varepsilon^{-1} W)^{-1}\right\}. & \end{array} \tag{4.9}$$

Model (4.9) can be written in state-space form so that the loglikelihood of $y^* = (y_1^*, \ldots, y_T^*)$ can be computed using (4.6).

The state vector α_t and the measurement matrix Z_t for model (4.8) do not involve μ:

$$\begin{array}{rcl} \alpha_t & = & \left(\begin{array}{ccccccccc} \mu_t^* & \nu_t^* & \beta_t^{*1} & \ldots & \beta_t^{*K} \end{array} \right)' \\ Z_t & = & \left(\begin{array}{ccccccccc} W & 0 & W x_t^{*1} & \ldots & W x_t^{*K_1} & W X_t^{*K_1+1} & \ldots & W X_t^{*K} \end{array} \right) \end{array} \tag{4.10}$$

The measurement equation for y_t^* is:

$$y_t^* = \mu^* + P' Z_t \alpha_t + u_t \tag{4.11}$$

where

$$P' Z_t = \left(\begin{array}{ccccccc} I & 0 & x_t^{*1} & \ldots & x_t^{*K_1} & X_t^{*1} & \ldots & X_t^{*K} \end{array} \right). \tag{4.12}$$

The measurement equation for y_t in model (4.1) is:

$$y_t = \mu + Z_t \alpha_t + \varepsilon_t. \tag{4.13}$$

Following Jungbacker & Koopman (2008), the loglikelihood $l(y^*)$ can be adjusted to calculate the non-diffuse likelihood of y in model (4.1) as follows:

$$l(y) = -\frac{(S-R)}{2}\left(T\log(2\pi) + \log(T)\right) + l(y^*) - \frac{T-1}{2}\log\frac{|\Sigma_\varepsilon|}{|\,(W'\Sigma_\varepsilon^{-1}W)^{-1}\,|} - \frac{1}{2}\sum_{t=1}^{T}\tilde{y}_t'\Sigma_\varepsilon^{-1}\tilde{y}_t, \tag{4.14}$$

where $\tilde{y}_t = (I - WP')(y_t - \bar{y})$.

Jungbacker & Koopman (2008) derive a similar adjustment of the diffuse likelihood for y, given the diffuse likelihood for y^*.

Each evaluation of the likelihood therefore requires three steps:

- Compute y_t^* and each X_t^{*k} for the current value $\hat{\Sigma}_\varepsilon$,
- Compute the likelihood for y_t^* using (4.6),
- Compute $\tilde{y}_t$ and adjusted likelihood for y_t with (4.14).

The number of knots and their position have been chosen a priori so that the matrix W is constant during parameter estimation. The score vector is computed numerically from the adjusted likelihood. Although the EM algorithm can be applied, we did not use it in our empirical study.

Initialization of the state vector α_t: We use diffuse initialization for the constant term μ^* in the state vector α_t for model (4.9). The local linear trend μ_t^* is initialized as a fixed vector zero. The slope term ν_1^* as well as all stochastic (constant) regression coefficients $\beta_t^{*k}, k = 1, \ldots, K$, have diffuse initial conditions.

Estimation of the constant level vector μ: The estimation of μ is not required for the likelihood evaluation in (4.14), only $\mu^* = P'\mu$ is involved. However, for proper signal extraction for the original time series y_t, we use the formula of Jungbacker & Koopman (2008) for the minimum mean squared linear estimate (MMSLE) of μ, $\hat{\mu}$, and the corresponding variance, given $\hat{\mu}^*$ and $\mathrm{Var}(\hat{\mu}^*)$:

$$\begin{aligned} \hat{\mu} &= (I - WP')\bar{y} + W\hat{\mu}^*, \\ \mathrm{Var}(\hat{\mu}) &= (I - WP')\tfrac{1}{T}\Sigma_\varepsilon(I - WP')' + W\mathrm{Var}(\hat{\mu}^*)W', \end{aligned} \tag{4.15}$$

where $\hat{\mu}^*$ is the ML estimate of μ^* and $\mathrm{Var}(\hat{\mu}^*)$ the corresponding estimated variance.

4.3 Application to French hourly electricity load

In section 4.2, we described a smoothing spline dynamic regression model and two methods for parameter estimation. In this section we illustrate the methodology in a model for hourly electricity loads. We first present the dataset and the effective model. Then we describe our benchmark models and we give practical implementation details.

4.3.1 Data description

We apply model (4.1)-(4.2)-(4.3)-(4.4) in state-space representation (4.5)-(4.7) to French national hourly electricity loads, measured in MegaWatts (MW), for the ten years 1997-2007. It is well-known that these high-frequency data are subject to several time-series effects: long-term positive trend, yearly, weekly and daily seasonal patterns. Electricity demand also depends on exogenous variables related to weather conditions (especially in France where electric heating plays an important part in electricity demand in winter): data such as temperature, cloud-cover, humidity or wind-speed can be involved. The inclusion of these effects in an electricity load model depends on the country of the data, on the availability and quality of data as well as forecasts. Another difficulty with electricity data is the recurring effect of special days and bank holidays: most of these days occur on a different day of the week from one year to another and can induce long weekends with extra free bridge days.

The model is estimated on the sample period January 1^{st}, 1997 until August 31^{st}, 2006. The post-sample period, September 1^{st}, 2006 until August 31^{st}, 2007, is used for forecasting accuracy evaluation. Figure 4.1 shows some aspects of the dataset: panel (a) draws the daily mean of vector of hourly loads y_t for the whole dataset (sample and post-sample) to show the positive trend and repeating yearly pattern over the years and panel (b) draws only year 2006 to zoom in on the intrayearly pattern of electricity load. There are no missing data. However, for simplicity, special days are explicitly treated as missing and therefore excluded from the analysis:

- Bank holidays and related bridge days,
- Christmas/New Year period: from December 23^{rd} until January 3^{rd},
- Daylight saving days: last Sundays of March and October,
- Peak Day Withdrawal days (EJP in French).

Approximately 3100 days remain in the sample, with about 75000 hourly observations.

In our model, we use explanatory variables with different frequencies. Our scalar daily explanatory variables capture the recurring intra-yearly and intra-weekly patterns:

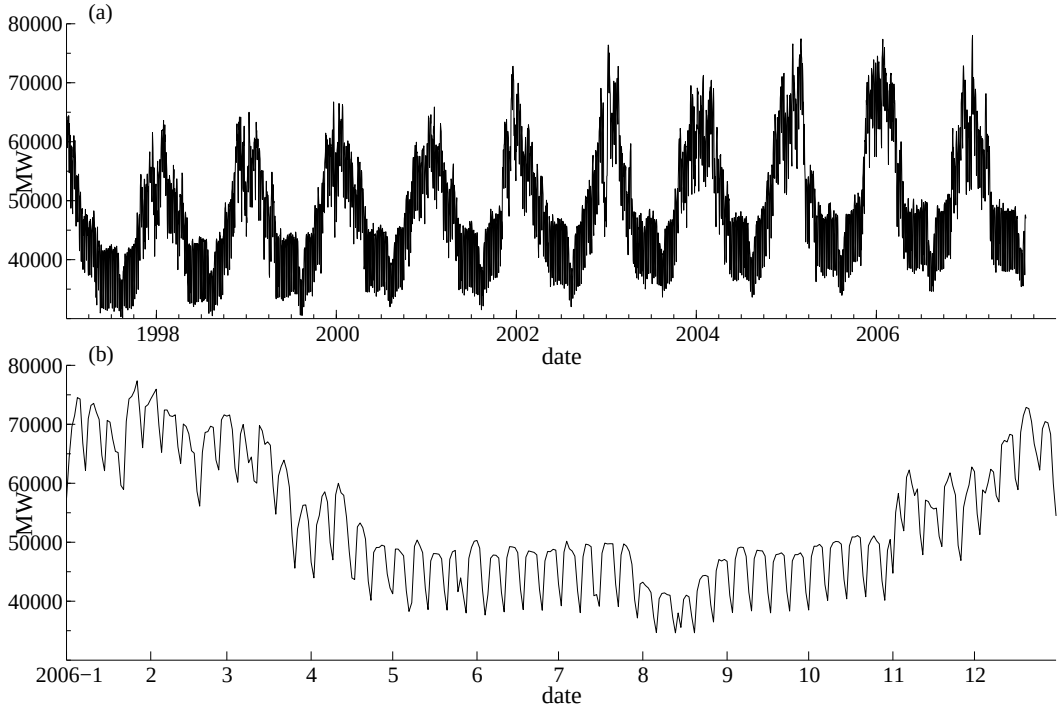

Figure 4.1: Daily mean of hourly electricity demand (a)from January 1^{st}, 1997 until August 31^{st}, 2007; (b) during year 2006.

- For the yearly patterns, we use Fourier coefficients as regressors

$$\begin{aligned} &x_t^1 = a_{1,t},\ x_t^2 = b_{1,t},\ x_t^3 = a_{2,t},\ x_t^4 = b_{2,t}, \\ &a_{i,t} = \cos\left(\tau_t \tfrac{2\pi i}{365.25}\right), \quad b_{i,t} = \sin\left(\tau_t \tfrac{2\pi i}{365.25}\right), \quad i = 1, 2, \end{aligned} \tag{4.16}$$

  where τ_t is the number of days elapsed since the 1^{st} of January in the year in which day t falls for $t = 1, \ldots, T$.

- For the weekly effects, we define 5 daytypes for all days of the week. The default daytype corresponds to regular Tuesdays, Wednesdays and Thursdays. We model electricity demand level differences with other daytypes using dummy variables, $x_t^5, x_t^6, x_t^7, x_t^8$, which respectively correspond to Mondays, Fridays, Saturdays and Sundays.

Our 24×1 vector hourly explanatory variables capture weather effects:

- For the heating effect, we define heating degrees and smoothed-heating degrees variables based on a national temperature $T_{s,t}$, a weighted average of local measures in ^oC:

$$\begin{aligned} &X_t^9 = \left(x_{0,t}^9 \ldots x_{23,t}^9\right)', \quad x_{s,t}^9 = \max\left(0, 15 - T_{s,t}\right), \\ &X_t^{10} = \left(x_{0,t}^{10} \ldots x_{23,t}^{10}\right)', \quad x_{s,t}^{10} = \max\left(0, 15 - T_{s,t}^{smo}\right), \end{aligned} \qquad s{=}0,\ldots,23, \tag{4.17}$$

where $T^{smo}_{s,t}$ is an exponentially smoothed national temperature ($\kappa = 0.98$ fixed a priori):

$$\begin{aligned} T^{smo}_{t,s+1} &= \kappa T^{smo}_{t,s} + (1-\kappa) T_{t,s+1}, \quad s = 1, \ldots, S-1, \\ T^{smo}_{t+1,1} &= \kappa T^{smo}_{t,S} + (1-\kappa) T_{t+1,1}, \quad s = S. \end{aligned} \tag{4.18}$$

- For the cooling effect, we define a smoothed-cooling degrees variable based on the smoothed national temperature variable $T^{smo}_{s,t}$:

$$X^{11}_t = \left(x^{11}_{0,t} \ldots x^{11}_{23,t}\right)', \quad x^{11}_{s,t} = \max\left(0, T^{smo}_{s,t} - 18\right), \quad s = 0, \ldots, 23. \tag{4.19}$$

- For the cloud-cover effect, we use $C_{s,t}$, a measure of national cloud-cover (without unit):

$$X^{12}_t = \left(x^{12}_{0,t} \ldots x^{12}_{23,t}\right)', \quad x^{12}_{s,t} = C_{s,t} \mathbf{1}_{T_{s,t}<15}, \quad s = 0, \ldots, 23. \tag{4.20}$$

We therefore consider $K = 12$ explanatory variables in equation (4.2) of model (4.1)-(4.2)-(4.3)-(4.4).

4.3.2 Model and benchmarks

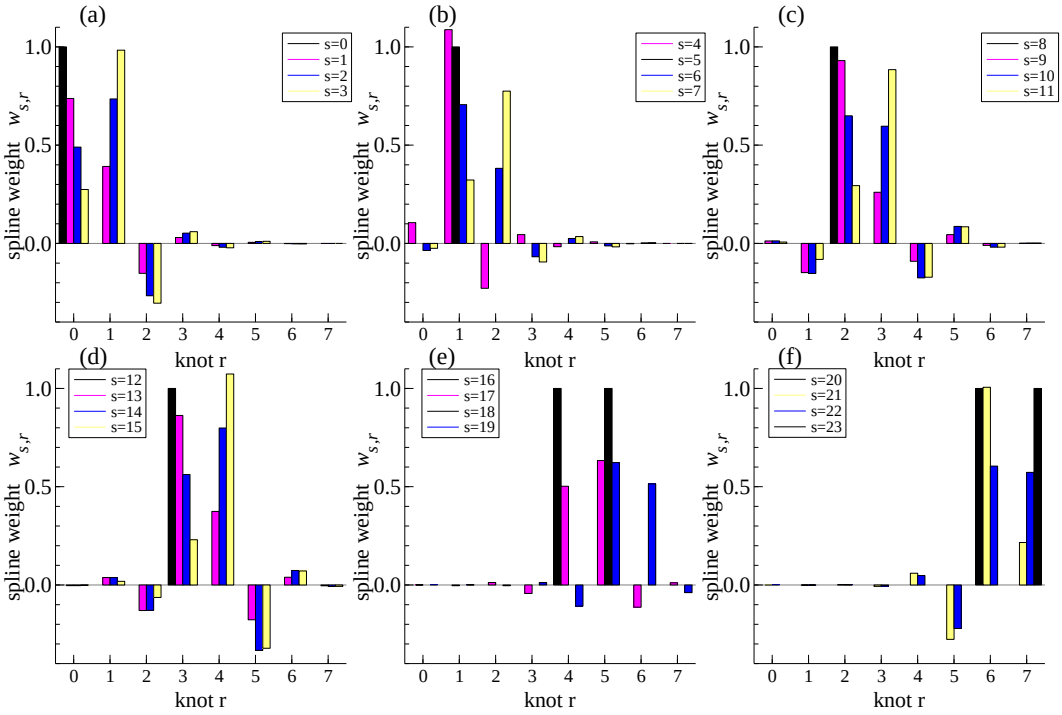

Figure 4.2: Spline weights for all hours (black unit bars correspond to knots): (a) $s = 0, 1, 2, 3$; (b) $s = 4, 5, 6, 7$; (c) $s = 8, 9, 10, 11$; (d) $s = 12, 13, 14, 15$; (e) $s = 16, 17, 18, 19$; (f) $s = 20, 21, 22, 23$. Knots at 0,5,8,12,16,18,20,23.

Define the load at day t and hour s by $y_{s,t}, s = 0, \ldots, S-1, t = 1, \ldots, T$, where $S = 24$ and T corresponds to August 31^{st}, 2006. Forecasts correspond to dates $t =$

$T+1,\ldots,T+365$. The $S\times 1$ vector $y_{s,t}$ represents the so-called daily load curve and is modelled by (4.1). The intradaily pattern is modelled by the smoothing spline. We set $R=8$ and the knot positions corresponding to hours $s^{\dagger}=\{0,5,8,12,16,18,20,23\}$, so that the spline weights matrix W is fixed. The setting is arbitrary, however we chose to make a balance between equally spaced knots and putting more knots for peak hours when there are quick changes in the daily load curve. Equation (4.1) is then fully specified. For illustration we present in Figure 4.2 the spline weights, $w_{s,r}$, for each knot $r=0,\ldots,R-1$ and each hour s in the model. Each hour is a weighted average of up to 6 knots, where weights sum to one.

The dynamic regression model for the knots, based on equation (4.2), is written as follows:

$$\lambda_t=\mu_t^*+\sum_{k=1}^{8}x_t^{*k}\beta_t^{*k}+\sum_{k=9}^{11}X_t^{*k}\beta_t^{*k}+X_t^{*12}\beta^{*12},t=1,\ldots,T, \tag{4.21}$$

where μ_t^* is the stochastic trend as in (4.3). The $R\times 1$ vectors of stochastic regression coefficients $\beta_t^{*k},k=1,\ldots,11$ follow multivariate random walks as in equation (4.4) and correspond to the scalar and vector explanatory variables $x_t^{*k},k=1,\ldots,8$ and $X_t^{*k},k=9,\ldots,11$. Finally, β_t^{*12} is the $R\times 1$ constant vector of regression coefficients associated with the cloud-cover related explanatory variable in X_t^{*12}.

Remark We impose smooth time trends by fixing all diagonal elements in Σ_η and Σ_ζ to 5. Not restricting these values can make the trends capture the intrayearly patterns, which we want to avoid.

- We denote by model A the electricity load model based on equations (4.1)-(4.21)-(4.3)-(4.4). We are interested in the comparison with the benchmark of models B and C, which are described below.

- Model B corresponds to the block diagonal dynamic factor time-varying regression model as described in chapter 3 and in Dordonnat et al. (2009). This model combines dynamics from three univariate hourly models in scalar dynamic factors. Here we consider the same empirical specification for electricity load (8 trivariate models and the same explanatory variables), except for the extra regression effects for morning hours 6 to 8. In comparison with model B, the positive side of model A is that it models all hours together reducing the number of parameters to estimate. However, model A also implies that the factor loadings are imposed a priori once R and the knot positions are fixed and that the factor loadings are the same for all dynamic components in the model. Some flexibility can be introduced in model A

by considering different spline weight matrices (in size R and/or knots position) for different dynamic components. We leave this for further empirical research. The results of Jungbacker & Koopman (2008) imply that (S-R) linear combinations of hourly loads follow white noise processes. This restriction is not supported by the data.

- Model C corresponds to the Univariate (or multivariate independent) modelling of each $y_{s,t}, s = 1, \ldots, S$. Each univariate model is based on the same dynamic components as model A. Model C therefore corresponds to model A with $W = I$ so that hourly explanatory variables are only transformed into diagonal matrices $X_t^{*k} = \text{diag}(X_t^k)$, daily explanatory variables remain unchanged. The S models can be estimated independently since all covariance matrices in (4.1)-(4.3)-(4.4) are diagonal. However, without dimension reduction in the dynamics, the total number of parameters to estimate is larger than for model A.

Models B and C are also applied to the dataset described in 4.3.1 and the results are compared with model A in section 4.4.

4.3.3 Practical implementation

Models A, B and C are implemented using `Ox`, see Doornik (2006), an object-oriented matrix programming environment. Efficient state-space modelling is performed using the `SsfPack` package of Koopman et al. (1999) and Koopman et al. (2008). This package provides state-space routines such as Kalman filtering, smoothing and likelihood evaluation for multivariate state space models. Exact treatment of diffuse initial conditions is included. We use these routines to estimate unknown parameters by maximizing the (diffuse) loglikelihood function. Then we can perform signal extraction and we compute forecasts to evaluate the three models, in-sample and out-of-sample.

For efficient filtering and smoothing, we follow Koopman & Durbin (2000) by including the irregular term u_t from equation (4.11) with non-diagonal covariance matrix in the state vector α_t. The resulting state-space form has no other irregular term in the measurement equation. This is computationally convenient. Models B and C are directly in the right form for `SsfPack`.

The remaining vector of parameters to estimate in model A is:

$$\psi = \left((\sigma_{\varepsilon,s})_{s=0,\ldots,23}, (\sigma_{k,r})_{k=1,\ldots,11;r=0,\ldots,7} \right)$$

We perform unconstrained optimisation by maximizing the (diffuse) loglikelihood function with respect to $\ln(\psi)$. For our hourly electricity load application, we estimate 144

parameters, 8 of which are treated as state variables (β_t^{*12} in (4.21)), and including the 24×1 vector μ estimated as in (4.15), see Tables 4.1 and 4.2 below, while we estimate a total of 504 parameters (8 independent estimations of 63 parameters), of which 184 coefficients are treated as state variables, for model B and 312 parameters for model C (24 univariate models with 13 parameters), of which 24 coefficients are treated as state variables. Note that we fix the variance at 5 for the trend component (both level and slope) for all three models, see the remark in section 4.3.2.

4.4 Results

4.4.1 Estimation results

Model (4.1)-(4.21)-(4.3)-(4.4) has been implemented and unknown parameters were estimated by maximizing the (diffuse) loglikelihood function using the method described in section 4.2.2. We also implemented the state-space model using the general method described in section 4.2.2 to compare likelihood evaluation computation time. The method of Jungbacker & Koopman (2008) reduced computation time by a factor three so that model estimation is manageable even with a long dataset and many explanatory variables. Good starting values for the likelihood maximization can be obtained e.g. from the univariate model C.

Table 4.1: Estimation results for model A on in-sample period January 1, 1997, until August 31, 2003.
Model A is defined in § 4.3.2. Remaining results are in Table 4.2, standard errors are in parentheses.

Par.	Est.	Par.	Est.	Par.	Est.	Par.	Est.
$\sigma_{\varepsilon,0}$	551	μ_0	22178(1129)	$\sigma_{\varepsilon,12}$	947	μ_{12}	31100(973)
$\sigma_{\varepsilon,1}$	0.14	μ_1	28193(2217)	$\sigma_{\varepsilon,13}$	25	μ_{13}	30621(952)
$\sigma_{\varepsilon,2}$	0.21	μ_2	31445(3897)	$\sigma_{\varepsilon,14}$	0.78	μ_{14}	29392(978)
$\sigma_{\varepsilon,3}$	384	μ_3	33405(5162)	$\sigma_{\varepsilon,15}$	297	μ_{15}	28115(1018)
$\sigma_{\varepsilon,4}$	600	μ_4	34787(5693)	$\sigma_{\varepsilon,16}$	479	μ_{16}	27400(864)
$\sigma_{\varepsilon,5}$	651	μ_5	35383(5228)	$\sigma_{\varepsilon,17}$	884	μ_{17}	28356(787)
$\sigma_{\varepsilon,6}$	1889	μ_6	34692(3720)	$\sigma_{\varepsilon,18}$	995	μ_{18}	30608(1023)
$\sigma_{\varepsilon,7}$	2323	μ_7	32572(1919)	$\sigma_{\varepsilon,19}$	1247	μ_{19}	29778(783)
$\sigma_{\varepsilon,8}$	1021	μ_8	29731(1171)	$\sigma_{\varepsilon,20}$	42	μ_{20}	28046(860)
$\sigma_{\varepsilon,9}$	381	μ_9	28597(1364)	$\sigma_{\varepsilon,21}$	1368	μ_{21}	26020(954)
$\sigma_{\varepsilon,10}$	605	μ_{10}	28555(1259)	$\sigma_{\varepsilon,22}$	1166	μ_{22}	27903(939)
$\sigma_{\varepsilon,11}$	684	μ_{11}	29706(1034)	$\sigma_{\varepsilon,23}$	0.12	μ_{23}	26217(1303)

Tables 4.1 and 4.2 detail estimated parameters. Table 4.1 gives estimates of mea-

surement error standard-deviations $\sigma_{\varepsilon,r}, r = 0, \ldots, 23$ in the vector of hyperparameters ψ as well as estimates based on $\hat{\psi}$ for the constant level $\mu_r, r = 0, \ldots, 23$ and associated standard-errors in parentheses following (4.15). The range of values for $\sigma_{\varepsilon,r}$ is wide, from almost 0 up to more than 2300. Standard-deviations are particularly large for hours 6,7,19 and 21 which don't have their own knot. The estimate of vector μ contains values that vary from 22.000 up to 35.000.

Table 4.2: Estimation results for model A on in-sample period January 1, 1997, until August 31, 2003.
Model A is defined in §4.3.2, see also Table 4.1. $s_r^\dagger$: knot positions - hour of the day.

	$s_r^\dagger$	0	5	8	12	16	18	20	23
Component	Param.	$r=0$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
Heating	$\sigma_{1,r}$	1.1E-03	2.6E-03	30.8	1.1E-03	2.6E-03	29.8	2.6E-03	35.6
Smoothed-heating	$\sigma_{2,r}$	1.1E-03	4.1E-03	1.6E-03	1.1E-03	4.1E-03	1.6E-03	4.1E-03	1.6E-03
Cooling	$\sigma_{3,r}$	1.6	1.8	7.7	1.6	1.8	7.5	1.7	7.5
Monday	$\sigma_{4,r}$	4.2E-03	1.9E-01	2.4E-01	4.2E-03	1.9E-01	2.4E-01	1.9E-01	2.4E-01
Friday	$\sigma_{5,r}$	4.8E-02	2.9E-01	2.9E-02	4.8E-02	2.9E-01	2.9E-02	2.9E-01	2.9E-02
Saturday	$\sigma_{6,r}$	81.0	78.3	158.9	128.0	151.6	138.6	113.4	96.6
Sunday	$\sigma_{7,r}$	93.0	114.3	228.2	186.0	205.2	156.3	132.1	93.8
a_1	$\sigma_{8,r}$	828.8	680.1	394.4	718.1	601.0	297.9	636.5	464.2
b_1	$\sigma_{9,r}$	994.5	808.9	408.8	758.5	703.1	419.2	664.0	470.0
a_2	$\sigma_{10,r}$	6.9	5.0	3.9	6.9	5.0	3.9	5.0	3.9
b_2	$\sigma_{11,r}$	4.6E-02	6.4	1.0E-04	4.6E-02	6.4	1.0E-04	6.4	1.0E-04
Cloud-cover	β^{*12}	6.0	-40.5	92.9	156.7	190.0	115.6	98.3	45.1
		(11.54)	(9.99)	(9.84)	(9.81)	(9.40)	(9.86)	(9.11)	(7.81)

Table 4.2 provides estimates for the dynamic regression component standard deviations $\sigma_{k,r}$, $k = 1, \ldots, 11$, $r = 0, \ldots, 7$ and for the constant regression coefficient vector for the cloud-cover explanatory variable $\hat{\beta}^{*12}$. The heating effect variance varies between the knots: only three coefficients are estimated as time-varying with standard-deviations away from zero. For the cooling effect, standard-deviations are small but sufficiently large to exhibit smooth variation of the coefficients over time. Standard-deviation estimates for smoothed-heating, for Mondays' and Fridays' daytype coefficients are almost zero. Model A estimated on these data does not exhibit time-variation for those coefficients. Regarding weekend daytype standard-deviations, estimates are larger for knots between 7 in the morning and 7 in the evening showing more time-variation for these hours. Finally, standard-deviations associated with Fourier coefficients are highly time-varying for the annual frequency while standard-deviations are much smaller for the second frequency.

The cloud-cover coefficient is significant for all knots except the first one at midnight. The main effect of this explanatory variable occurs during day hour related knots.

4.4.2 Signal extraction

Using matrix W and the maximum likelihood estimate of Σ_ε in Table 4.1, we transform the series of observations y_t into the lower dimension vector y_t^* using (4.8). Figure 4.3 draws each element of vector y_t^*. The main features of hourly electricity demand remain visible.

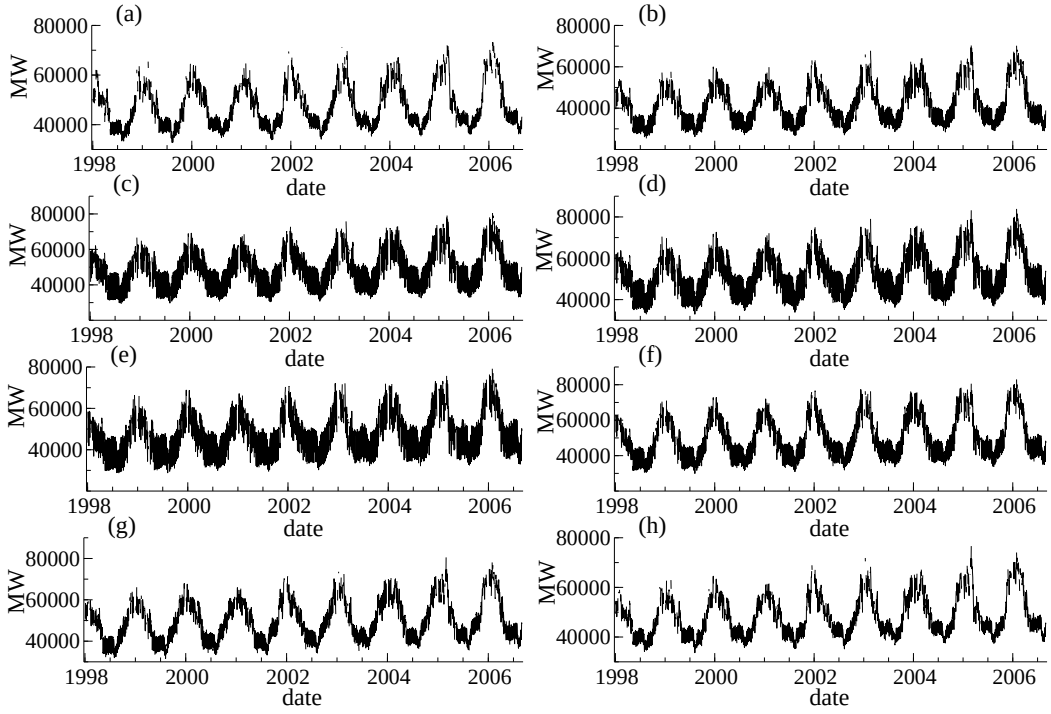

Figure 4.3: Time series of each element in knot vector y_t^* from model A defined in §4.3.2 using (4.8): (a) $y_{0,t}^*$; (b) $y_{1,t}^*$; (c) $y_{2,t}^*$; (d) $y_{3,t}^*$; (e) $y_{4,t}^*$; (f) $y_{5,t}^*$; (g) $y_{6,t}^*$; (h) $y_{7,t}^*$.

Smoothed estimates of the state vector $\hat{\alpha}_t$

Using maximum likelihood estimates from Σ_ε in Table 4.1 and dynamic component standard-deviations for stochastic regression effects in Table 4.2, the Kalman smoothing algorithm outputs $\hat{\alpha}_t$, the estimate of the state vector α_t based on all in-sample data. The algorithm can be applied to the original data vector y_t as well as on the lower dimensional data vector y_t^*, the latter method allowing faster computations. Although the sample starts on January 1^{st},1997, all graphs presented in this section start on January 1^{st},1998 after the initialization process for α_t has been completed.

Figure 4.4 exhibits the smoothed estimates of the stochastic regression coefficients associated with weather-related explanatory variables. Figure 4.4(a) draws in-sample estimates for the heating degrees regression coefficients $\hat{\beta}_{r,t}^{*9}, r = 0, \ldots, 7$. Five of them are actually estimated as constant regression coefficients while the three remaining ones

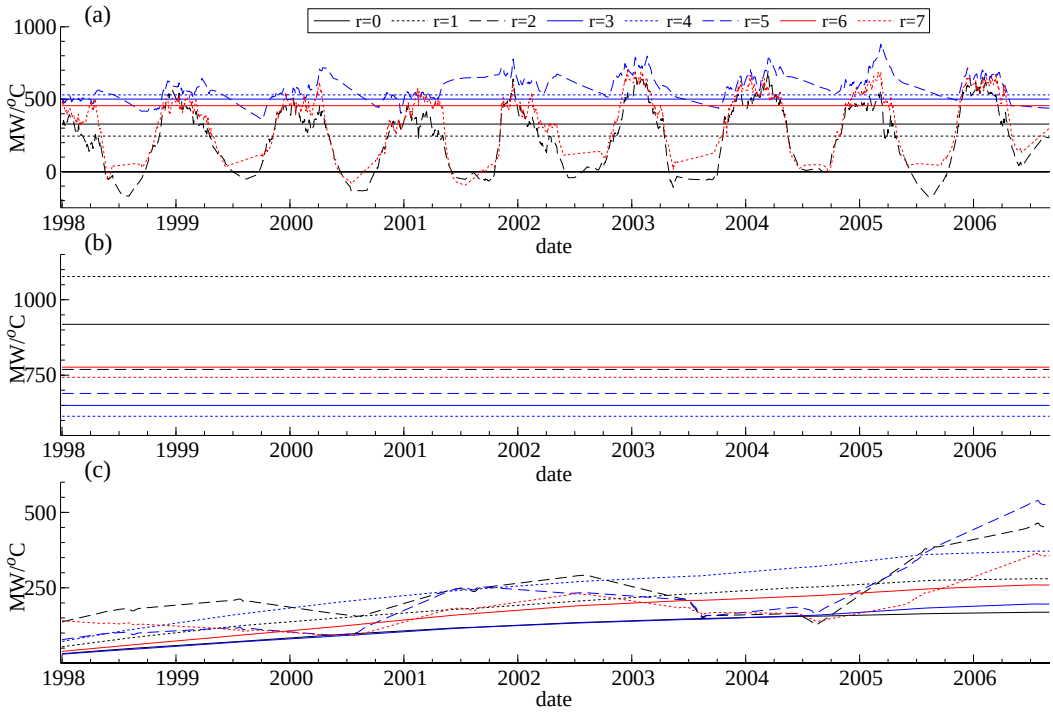

Figure 4.4: Estimated regression coefficients for weather-based explanatory variables for knot hours in model A defined in §4.3.2: (a) heating $\hat{\beta}^{*9}_{r,t}, r = 0, \ldots, 7$; (b) smoothed-heating $\hat{\beta}^{*10}_{r,t}, r = 0, \ldots, 7$; (c) cooling $\hat{\beta}^{*11}_{r,t}, r = 0, \ldots, 7$.

exhibit an intra-yearly pattern with a strong increase at the beginning of the heating period and a fast decrease at the end of the period. The coefficients vary between 250 and 750 MW/^{o}C. This first direct heating effect is augmented with the smoothed-heating degrees regression coefficients estimates $\hat{\beta}^{*10}_{r,t}, r = 0, \ldots, 7$ shown in Figure 4.4(b). For all knots, the regression coefficients are constant over time. The estimated values vary from 600 to almost 1100 MW/^{o}C. Finally, Figure 4.4(c) draws the estimated cooling degrees regression coefficients $\hat{\beta}^{*11}_{r,t}, r = 0, \ldots, 7$. All coefficients exhibit a positive trend from less than 200 MW/^{o}C in 1998 up to 500 MW/^{o}C for some knots at the end of the estimation sample. The weather-related effects on electricity loads are modelled in a quite different way compared to Dordonnat et al. (2009), as in chapter 3. The spline smoothing reduces the dimension of the hourly explanatory variables. The constant estimate of the smoothed-heating regression effect may be explained by this transformation. Since the explanatory variables themselves are transformed in our model, regression coefficients cannot be compared directly with estimates from benchmark models B and C. In this case we consider regression effects instead. This particular treatment is not required for daily explanatory variables. These allow comparison of the regression coefficients directly with benchmark models.

Figure 4.5 details the stochastic regression coefficient estimates associated with the

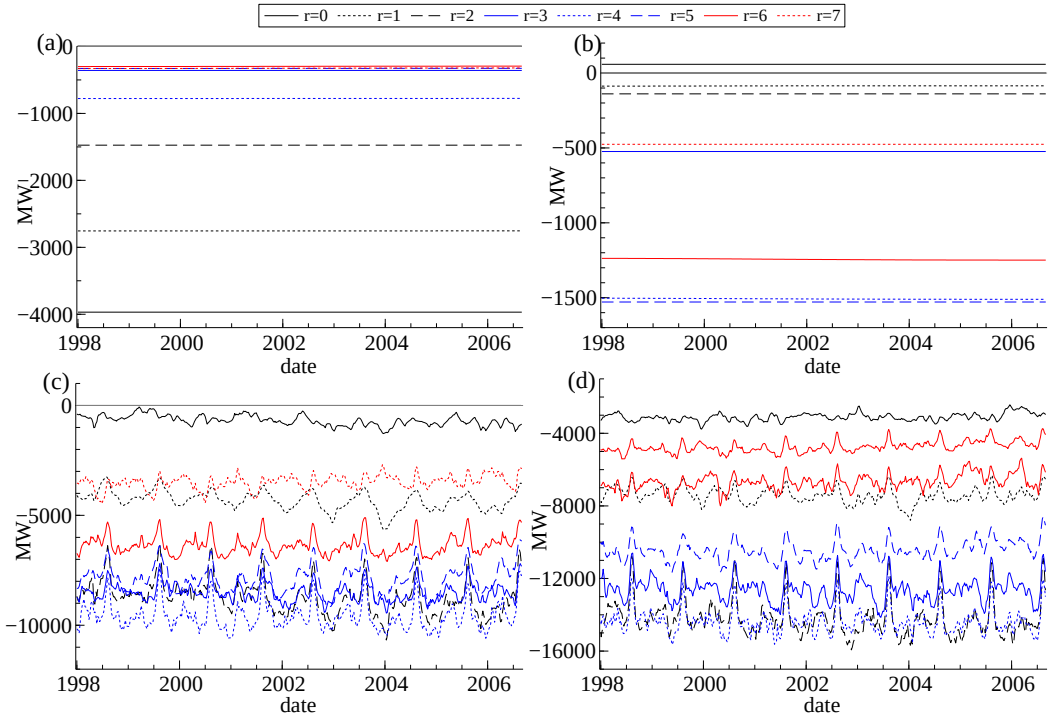

Figure 4.5: Estimated regression coefficients for daytype dummies for knot hours in model A defined in §4.3.2: (a) Mondays $\hat{\beta}^{*5}_{r,t}, r = 0, \ldots, 7$; (b) Fridays $\hat{\beta}^{*6}_{r,t}, r = 0, \ldots, 7$ (c) Saturdays $\hat{\beta}^{*7}_{r,t}, r = 0, \ldots, 7$ (d) Sundays $\hat{\beta}^{*8}_{r,t}, r = 0, \ldots, 7$.

daytype dummy explanatory variables. Figure 4.5(a) shows the load decrease in MW for Mondays $\hat{\beta}^{*5}_{r,t}, r = 0, \ldots, 7$: the effect is constant throughout the sample for all knots. The effect is highly negative in the morning (-4000 MW) and tends to 0 during the day. The weekend impacts electricity consumption on the Monday morning but the end of the day is more like a regular working day. We observe the opposite situation in Figure 4.5(b) which draws the regression coefficients for Fridays $\hat{\beta}^{*6}_{r,t}, r = 0, \ldots, 7$. The effect is also constant for the estimation period and close to 0 in the morning. A minimum is reached in the afternoon at -1500 MW for knots 5 and 6. The effect is less important for evening knots. Figure 4.5(c) and (d) finally exhibit coefficient estimates for Saturdays $\hat{\beta}^{*7}_{r,t}, r = 0, \ldots, 7$ and Sundays $\hat{\beta}^{*8}_{r,t}, r = 0, \ldots, 7$. The effect for both daytypes is far more important than the Monday and Friday effects. The weekend effect is more pronounced during day hours. It goes down to -10.000 MW for Saturdays and -15.000 MW for Sundays. For the day hour knots, we also notice repeating peaks in August that reduce the load level difference between weekdays and weekends. During this month, there is a global decrease in economic activity which reduces the weekend effect.

Smoothed dynamic components for y_t

We have shown the evolution of most components of the state vector α_t in the sample. Using the level estimate $\hat{\mu}$, the observation matrix Z_t defined in (4.10) based on spline weight matrix W and daily explanatory variables $x_t^{*k}, k = 1, \ldots, 8$ and transformed explanatory hourly variables $X_t^{*k}, k = 9, \ldots, 12$, we obtain estimated regression effects and associated standard-errors for each time series $y_{s,t}, s = 0, \ldots, 23$ using smoothed estimate $\hat{\alpha}_t$. Note that for daily explanatory variables, spline smoothing does not affect the variables themselves, only the stochastic regression coefficient. We draw smoothed estimates $W\hat{\beta}_t^{*k}, k = 1, \ldots, 8$ and compare them with corresponding stochastic regression coefficients estimated with models B and C. For the hourly explanatory variables this is not possible since the variables are transformed using the P' matrix. We can however compare estimated regression effects $WX_t^{*k}\hat{\beta}_t^{*k}, k = 9, \ldots, 12$ with the corresponding ones from models B and C.

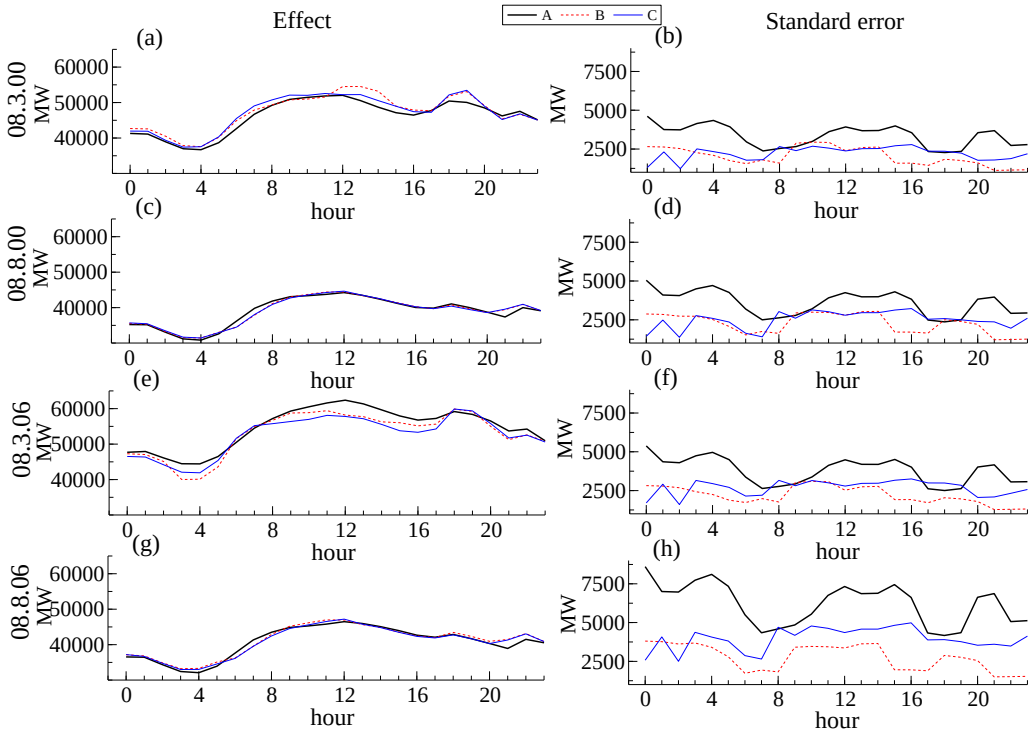

Figure 4.6: Estimated trend and yearly pattern component for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue): (a) estimated effect and (b) estimated standard-error on Wednesday March 8^{th}, 2000; (c) estimated effect and (d) estimated standard-error on Tuesday August 8^{th}, 2000; (e) estimated effect and (f) estimated standard-error on Wednesday March 8^{th}, 2006; (g) estimated effect and (h) estimated standard-error on Tuesday August 8^{th}, 2006.

Figure 4.6 draws the trend and yearly component ($\hat{\mu} + W\hat{\mu}_t^* + \sum_{k=1}^{4} WX_t^{*k}\hat{\beta}_t^{*k}$ for model A) for some specific dates: Figure 4.6 (a) corresponds to the estimated effect

on Wednesday March 8^{th}, 2000 and (b) draws the corresponding standard-errors for models A, B and C; (c) and (d) correspond to Tuesday August 8^{th}, 2000, (e) and (f) to Wednesday March 8^{th}, 2006 and finally (g) and (h) to Tuesday August 8^{th}, 2006. In all cases the three models give consistent estimates with a flatter daily curve in August. For March figures, the difference between models can reach 5000 MW. Regarding standard errors, model A gives larger values, while models B and C tend to give similar values except for Figure (h).

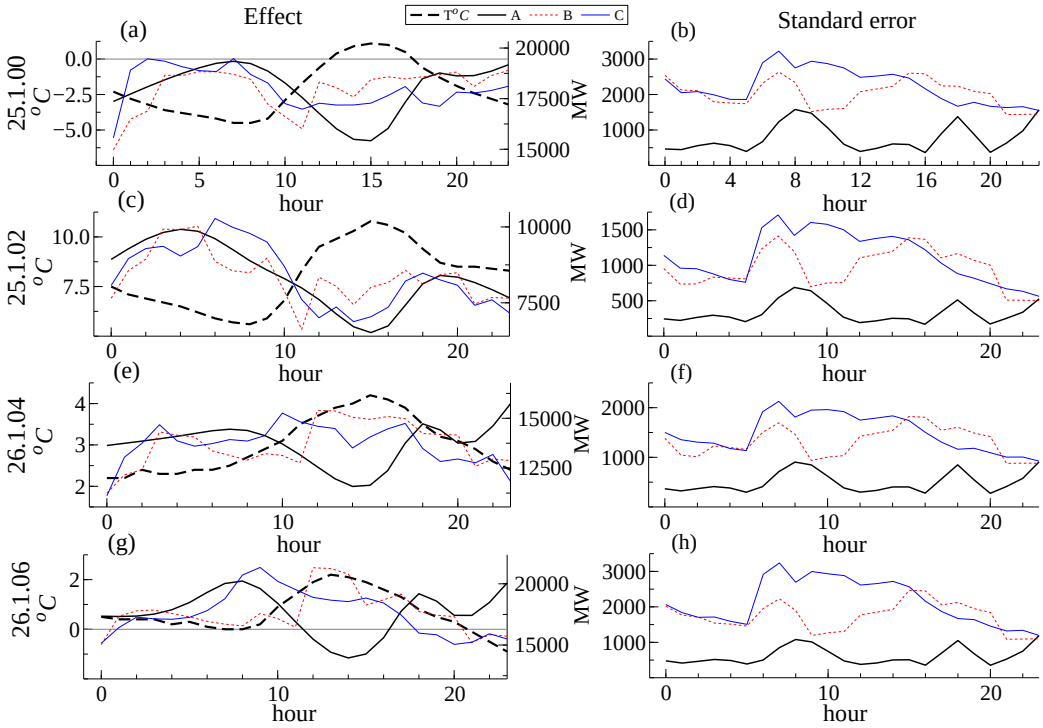

Figure 4.7: Estimated overall heating and smoothed-heating effect for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue), with observed national temperature (black bold dashed line)- left y-axis corresponds to the temperature in ^oC and right y-axis corresponds to the regression effect/standard-error in MW:
(a) estimated effect and (b) estimated standard-error on Tuesday January 25^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on Friday January 25^{th}, 2002;
(e) estimated effect and (f) estimated standard-error on Monday January 26^{th}, 2004;
(g) estimated effect and (h) estimated standard-error on Thursday January 26^{th}, 2006.

Figure 4.7 draws the estimated overall heating effect for models A ($WX_t^{*9}\hat{\beta}_t^{*9} + WX_t^{*10}\hat{\beta}_t^{*10}$), B and C on Tuesday January 25^{th}, 2000 (a) with associated standard-errors (b), on Friday January 25^{th}, 2002 (c) and (d), on Monday January 26^{th}, 2004 (e) and (f) and on Thursday January 26^{th}, 2006 (g) and (h). We also provide the daily pattern of national temperature on the corresponding days so we can see the opposite movement of temperature and heating effect. Compared to model B and C, model A estimates a smaller heating effect especially during the afternoon of cold days. However,

model A estimates follows a clear opposite movement compared to the temperature while models B and C estimates are counter-intuitively increasing with temperature, especially in the afternoon. Associated standard-errors for model A are smaller and smooth across day.

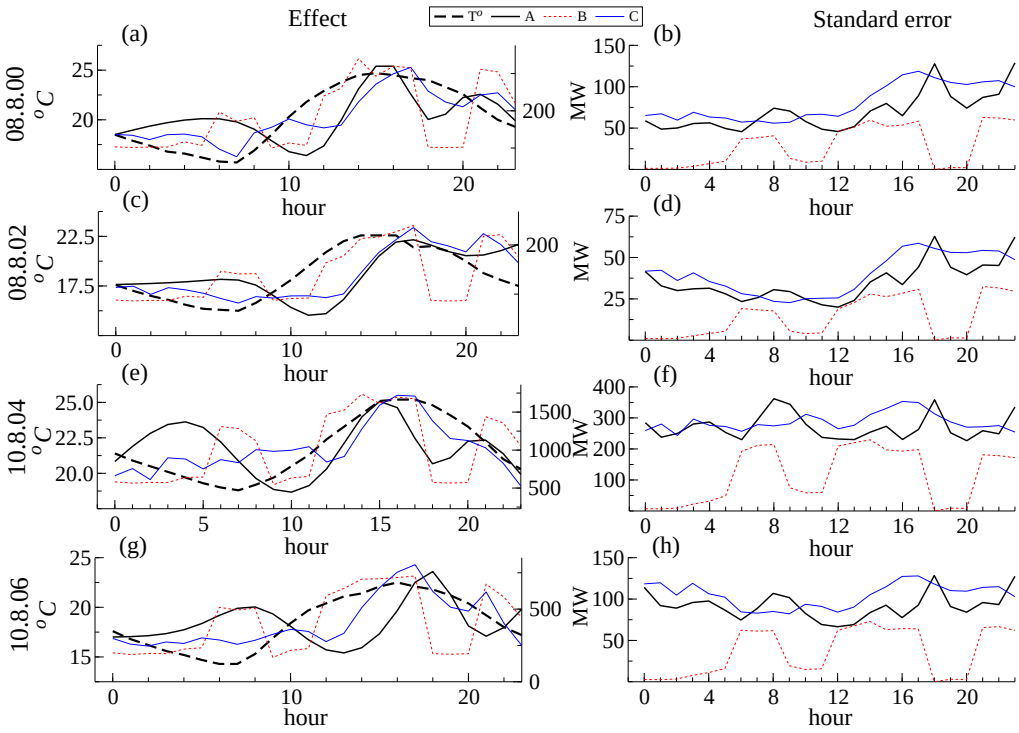

Figure 4.8: Estimated cooling effect for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue), with observed national temperature (black bold dashed line)- left y-axis corresponds to the temperature in ^oC and right y-axis corresponds to the regression effect/standard-error in MW:
(a) estimated effect and (b) estimated standard-error on Tuesday August 8^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on Thursday August 8^{th}, 2002;
(e) estimated effect and (f) estimated standard-error on Tuesday August 10^{th}, 2004;
(g) estimated effect and (h) estimated standard-error on Thursday August 10^{th}, 2006.

Figure 4.8 draws the estimated cooling effect for models A ($WX_t^{*11}\hat{\beta}_t^{*11}$), B and C on Tuesday August 8^{th}, 2000 (a) with associated standard-errors (b), on Thursday August 8^{th}, 2002 (c) and (d), on Tuesday August 10^{th}, 2004 (e) and (f) and on Thursday August 10^{th}, 2006 (g) and (h). Note that this feature is difficult to capture in a statistical model since most of sufficiently warm days to observe cooling occur in August when we also observe an opposite effect due to the lower economic activity in the summer. These two effects are hard to distinguish. In Figures (a) and (c) estimates based on model A and C are quite consistent while model A gives a more pronounced daily pattern of the cooling effect in Figure 4.8(g). Model B is unsatisfactory for this feature, estimating a fixed cooling effect for two groups of hours, (0-2) and (18-20). Standard errors for

model B are relatively small but this is due to misspecification of the model for this purpose. Models A and C give similar levels for their hourly standard-errors. The daily pattern of the national temperature is not clearly leading the cooling effect for the night hours: temperature variations do not affect cooling effect during the night. This can be explained by the main part of cooling demanded by offices, closed during the night. The daily pattern of the cooling effect seems too much pronounced for model A, especially the strong decrease at the end of the morning. Note that the overall cooling effect in the summer is much smaller (most of the time far less than 1.000 MW) than the heating effect in the winter.

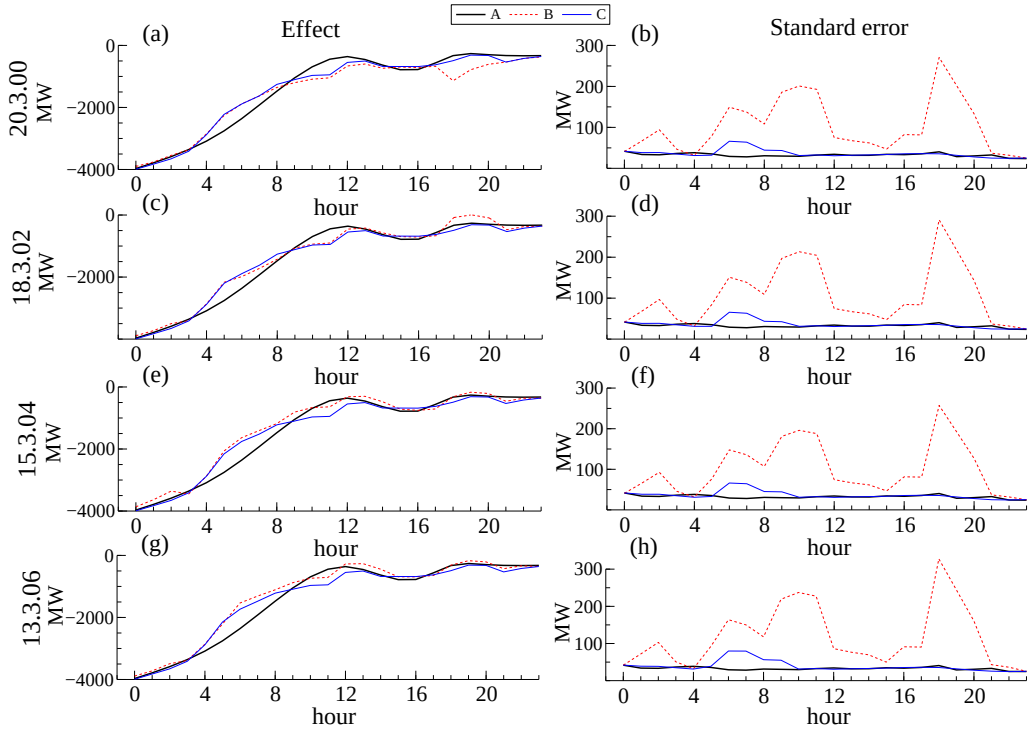

Figure 4.9: Estimated Monday effect for original time series y_t across hour of the day for models A(black bold line), B(red dashed) and C(thin blue):
(a) estimated effect and (b) estimated standard-error on March 20^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on March 18^{th}, 2002;
(e) estimated effect and (f) estimated standard-error on March 15^{th}, 2004;
(g) estimated effect and (h) estimated standard-error on March 13^{th}, 2006.

Figure 4.9 draws the estimated effect of Mondays for models A, B and C on March 20^{th}, 2000 (a) with associated standard-errors (b), on March 18^{th}, 2002 (c) and (d), on March 15^{th}, 2004 (e) and (f) and on March 13^{th}, 2006 (g) and (h). Model A, by construction, gives a smoother daily curve but all models give consistent estimates: the smoothing spline used in model A is not too restrictive in this respect. The Monday effect is minimum at midnight with -4000 MW and reaches zero from noon to hour 23. During the first half of the day, load consumption is affected by the weekend but not

anymore for the second half. Standard errors are larger for model B since for most hours, the regression coefficient is indeed time-varying while the model C estimates are constant except for the morning peak hours when standards errors get larger. In model A, the estimated effect is also constant giving the smallest standard deviations. Model A gives similar values as model C with smaller values for morning peak hours 6 to 9.

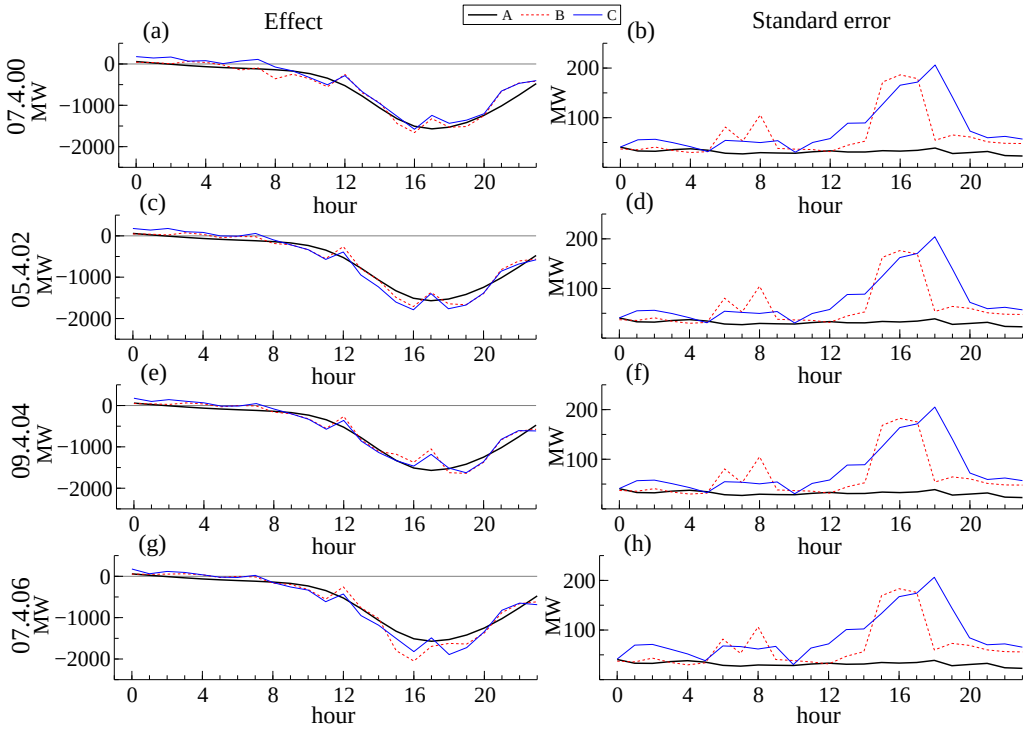

Figure 4.10: Estimated Friday effect for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue):
(a) estimated effect and (b) estimated standard-error on April 7^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on April 5^{th}, 2002;
(e) estimated effect and (f) estimated standard-error on April 9^{th}, 2004;
(g) estimated effect and (h) estimated standard-error on April 7^{th}, 2006.

Figure 4.10 draws the estimated effect of Fridays for models A, B and C on April 7^{th}, 2000 (a) with associated standard-errors (b), on April 5^{th}, 2002 (c) and (d), on April 9^{th}, 2004 (e) and (f) and on April 7^{th}, 2006 (g) and (h). Model A seems to smooth the daily curves estimated by models B and C. From midnight to 8 in the morning the effect is almost zero and is getting larger in absolute value from morning until reaching a minimum in the afternoon at hours 16-18 (around -1500 MW). Then the effect is less important in the evening to reach -500 MW. As for the Monday effect, standard-errors are larger for model B than for model A but this is also the case for model C. Standard-errors for model A are the same for all hours while models B and C give peaks for peak hours in the morning and evening.

Figure 4.11 draws the estimated effect of Saturdays for models A, B and C on March

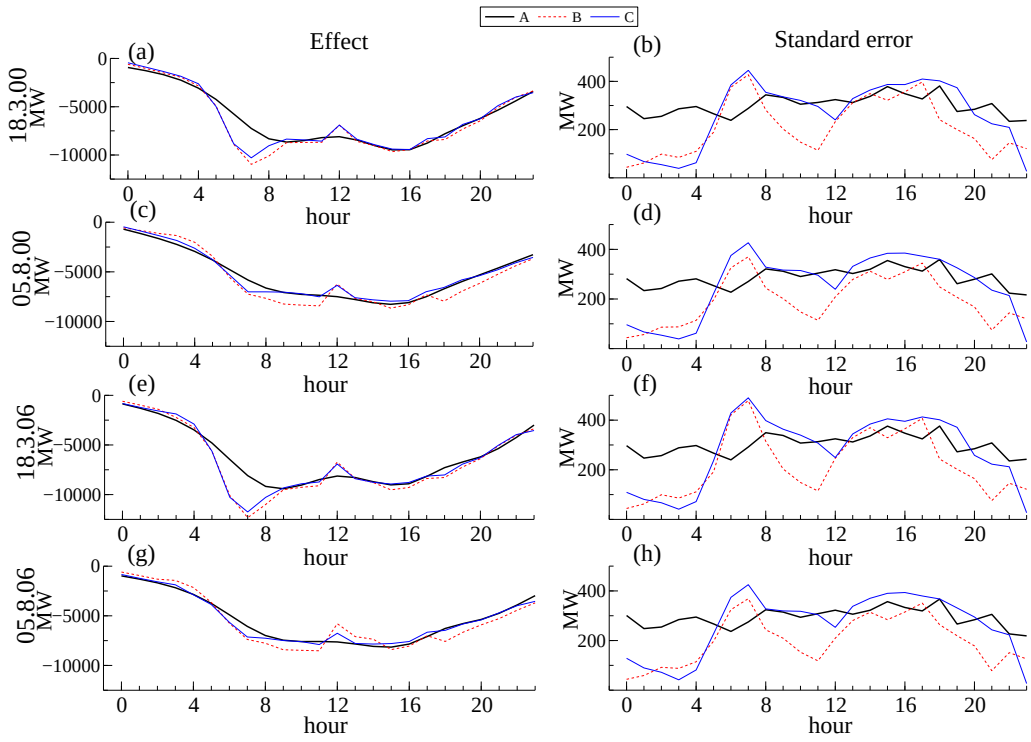

Figure 4.11: Estimated Saturday effect for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue):
(a) estimated effect and (b) estimated standard-error on March 18^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on August 5^{th}, 2000;
(e) estimated effect and (f) estimated standard-error on March 18^{th}, 2006;
(g) estimated effect and (h) estimated standard-error on August 5^{th}, 2006.

18^{th}, 2000 (a) with associated standard-errors (b), on August 5^{th}, 2000 (c) and (d), on March 18^{th}, 2006 (e) and (f) and on August 5^{th}, 2006 (g) and (h). Here we observe interesting differences between the three models. Models B and C are consistent with each other. For most hours, model A is also consistent with the two benchmark models. However in March we notice substantial differences in the morning hours 6 to 8 (the difference is about 3.400 MW at 6 and 4.000 MW at 7, that is about 5% of the load). Overall the three model have similar patterns: the Saturday effect is relatively small during the night but reaches large negative values during day hours (around -9.000 MW). Model A smoothes standard-errors as well but this time values are somewhat larger than for models B and C, especially in the morning hours up to 4.

Figure 4.12 draws the estimated effect of Sundays for models A, B and C on March 19^{th}, 2000 (a) with associated standard-errors (b), on August 6^{th}, 2000 (c) and (d), on March 19^{th}, 2006 (e) and (f) and on August 6^{th}, 2006 (g) and (h). In August, models A and C give similar daily patterns while in March models B and C are consistent with each other. Again model differences are substantial for the morning hours in March 2000 and 2006. Model B and C give results as expected. For this component, the estimated

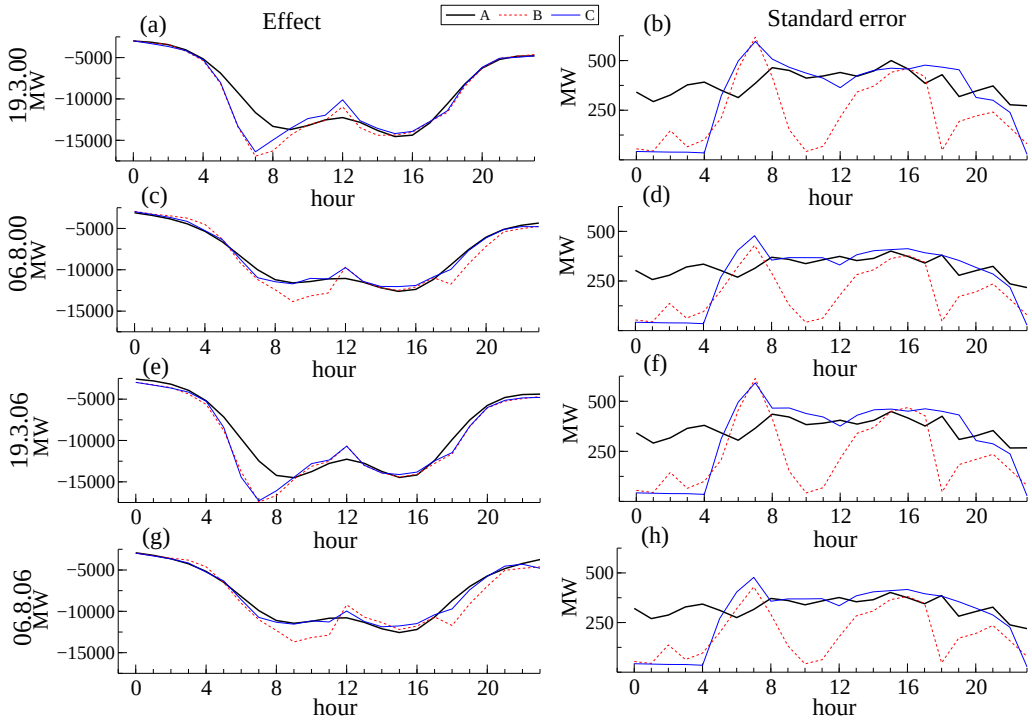

Figure 4.12: Estimated Sunday effect for original time series y_t across hour of the day for models A (black bold line), B(red dashed) and C(thin blue):
(a) estimated effect and (b) estimated standard-error on March 19^{th}, 2000;
(c) estimated effect and (d) estimated standard-error on August 6^{th}, 2000;
(e) estimated effect and (f) estimated standard-error on March 19^{th}, 2006;
(g) estimated effect and (h) estimated standard-error on August 6^{th}, 2006.

intradaily pattern could be improved by setting a knot at 7 instead of 8. The overall daily pattern estimated by the three models is similar to the Saturday pattern but estimated values are even larger at around -13.000 MW during day hours. While standard-errors have a pronounced daily pattern for model B, they are almost constant for model A and consistent with model C from 8 in the morning until hour 22, much larger for other hours.

Finally, Figures 4.13(a) and (c) show the estimated intraweekly pattern for y_t, i.e. the estimated trend $\hat{\mu} + W\hat{\mu}_t^*$ associated with all daytypes effects $\sum_{k=5}^{8} x_t^{*k} W \hat{\beta}_t^{*k}$, while figures 4.13(b) and (d) add to the previous the intrayearly pattern $\sum_{k=1}^{4} x_t^{*k} W \hat{\beta}_t^{*k}$ and give the estimated total weather-independent load. Figure 4.13 only concerns model A. Figures (a) and (b) draw 3 specific weeks in 2001 (one in April, one in the summer and one in November) while Figures (c) and (d) exhibit 3 specific weeks in 2005-2006 (one in November 2005, one in June 2006 and the last one during summer 2006). In all figures weekends are clearly distinguished from weekdays. Figures (b) and (d) show how the intraweekly pattern is evolving during a year, note especially the evening peak in November, which clearly increases over the years.

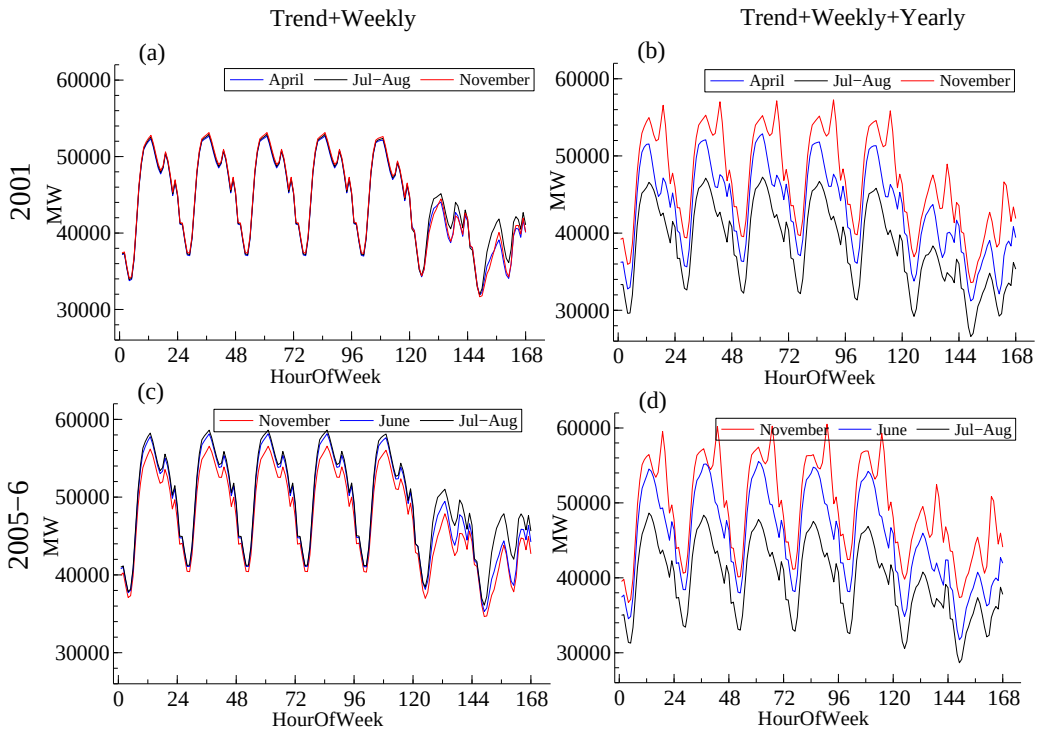

Figure 4.13: Estimated intraweekly pattern for original time series y_t for model A: trend and weekly pattern $\hat{\mu} + W\hat{\mu}_t^* + \sum_{k=5}^{8} x_t^{*k} W\hat{\beta}_t^{*k}$ (a) Apr. 23^{rd} - 29^{th}, 2001(blue line), Jul. 30^{th} - Aug. 5^{th}, 2001(black), Nov. 19^{th} - 25^{th}, 2001 (red);(c)Nov. 14^{th} - 20^{th}, 2005(red line), Jun. 12^{th} - 18^{th}, 2006(blue), Jul. 31^{st} - Aug. 6^{th}, 2006 (black); trend, weekly and yearly pattern $\hat{\mu} + W\hat{\mu}_t^* + \sum_{k=1}^{8} x_t^{*k} W\hat{\beta}_t^{*k}$ (b) Apr. 23^{rd} - 29^{th}, 2001(blue line), Jul. 30^{th} - Aug. 5^{th}, 2001(black), Nov. 19^{th} - 25^{th}, 2001 (red);(d)Nov. 14^{th} - 20^{th}, 2005(red line), Jun. 12^{th} - 18^{th}, 2006(blue), Jul. 31^{st} - Aug. 6^{th}, 2006 (black). Note : Weather effects are not included.

4.4.3 Model diagnostics

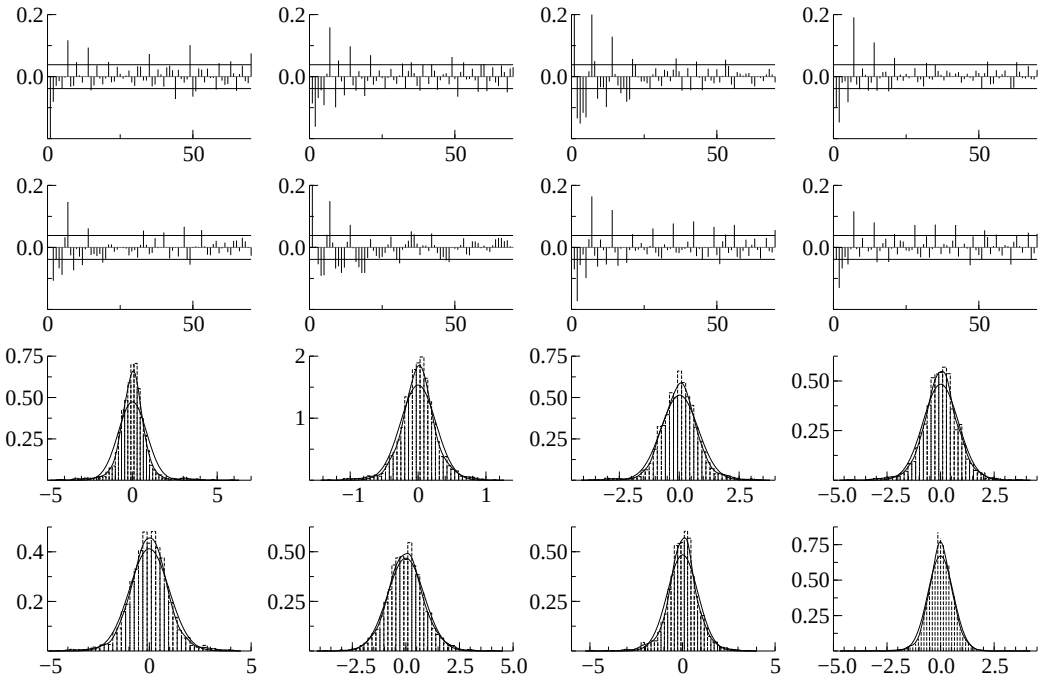

Figure 4.14: Estimated standardised residuals ACF (top), empirical distribution and Gaussian adjustments (bottom) for transformed time series y_t^*, row by row.

If the model is well specified the estimated in-sample standardised residuals produced by the Kalman filter should exhibit no dynamic structure and their distribution should be approximately Gaussian. Figure 4.14 shows the empirical ACF associated with each knot, as well as their empirical distribution, histogram and density, associated with the corresponding Gaussian distribution (same mean and standard-deviation as the empirical density). The empirical ACFs still exhibit some dynamic weekly structure. The empirical distributions exhibit most of the time good Gaussian adjustments. The first two knots have a larger kurtosis than the corresponding Gaussian densities. The diagnostics for the knots are therefore quite satisfactory. This is not the case for the standardised residuals for the original y_t, showing that there is still dynamic structure in the orthogonal complement of y_t^*, $P_\perp' y_t$ where $P_\perp$ is the $S \times (S - R)$ matrix with rank $S - R$ so that $P_\perp' P = 0$.

4.4.4 Forecasting results

Using the estimated state vector α_t for the last in-sample observation and explanatory variables transformed with the estimated P' matrix, the Kalman filter can be applied to obtain forecasts, from $T + 1$ up to $T + h$, with horizon $h > 1$, and treating the

corresponding days $T+1,\ldots,T+h$ as missing data. The issue of evaluating a model using realised or forecast values for the explanatory variables (as this would be done in a real forecasting process) is discussed in Dordonnat et al. (2008), see chapter 2. The relative forecasting accuracy of different models was not influenced by this issue. We only evaluate our model for one-step ahead forecasts only, as the likelihood maximization minimizes the standardised mean squared one-step ahead forecast errors by definition (4.6). Table 4.3 exhibits one-day ahead hourly forecasting accuracy on the post-sample period September 1^{st}, 2006 until August 31^{st}, 2007. As accuracy measures, we use the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE). We produce the results for model A and for our two benchmark models B and C.

Table 4.3: Hourly diagnostics - post-sample Root Mean Squared Error (RMSE) and MAPE (Mean Absolute Percentage Error) for one-day ahead forecasts for smooth dynamic factor model (A) and benchmark models, block dynamic factor model (B) and univariate model (C). N is the number of days actually forecasts in the post-sample.

		A		B		C	
Hour	N	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
0	307	1100	1.68	1013	1.33	937	1.29
1	307	956	1.36	973	1.33	919	1.29
2	307	919	1.45	950	1.39	872	1.35
3	307	938	1.60	787	1.23	818	1.26
4	307	961	1.72	722	1.16	745	1.22
5	307	1112	2.10	697	1.14	719	1.19
6	307	2347	4.27	897	1.44	895	1.45
7	307	2689	4.52	1047	1.53	1068	1.59
8	307	1465	2.29	859	1.18	840	1.14
9	307	1010	1.36	920	1.22	853	1.10
10	307	1071	1.46	950	1.24	931	1.17
11	307	1142	1.64	969	1.27	932	1.16
12	307	1399	2.01	847	1.12	823	1.05
13	307	1004	1.24	937	1.26	868	1.17
14	307	1079	1.37	994	1.35	923	1.27
15	307	1190	1.65	1053	1.43	971	1.36
16	307	1250	1.83	1092	1.49	999	1.45
17	307	1360	1.99	1024	1.41	959	1.37
18	307	1328	1.89	997	1.34	905	1.24
19	307	1458	2.08	939	1.24	852	1.14
20	307	811	1.18	819	1.14	764	1.12
21	307	1525	2.60	699	1.03	741	1.09
22	307	1332	2.22	684	0.96	696	0.97
23	307	693	1.02	662	0.94	682	0.97

Model A is clearly outperformed by both models B and C, which give similar forecasting accuracy. Forecasting is not the primary aim of this chapter, results are however

disappointing especially for morning and evening peak hours where forecasting accuracy is poor. Discussion is provided below.

Table 4.3 allows comparison between models A, B, and C but we also considered supplementary models for model A, varying the number and positions of knots:

- model D corresponds to the spline model A with $R = 11$ and knots at positions {0;3;6;7;8;12;15;17;19;21;23}; it gives far more satisfactory forecasting results than model A for the morning peak hours 6 to 8 and the late evening hours 21 to 23. However model A clearly outperforms model D for the evening peak hours 17 to 20. For the others hours, both models are comparable.
- model E corresponds to the spline model A with $R = 9$ and knots at positions {0;4;6;8;12;17;19;21;23}.
- model F corresponds to the spline model A with $R = 9$ and knots at positions {0;3;6;8;12;17;19;21;23}.

Models E and F show a similar pattern of forecasting accuracy as model D: better than A in the morning peak hours and worse than A in the late afternoon. They are closer to model A in the late evening. The daily shapes of the forecasting MAPE and RMSE are similar to the ones of the internal EDF model, which are also exacerbated for peak hours.

4.4.5 Discussion

In this section we want to underline the following points and put them in perspective:

- Estimation of effects for daily variables,
- The transformation of hourly explanatory variables related to weather conditions, and
- Improving the method to reach a better forecasting accuracy.

We are satisfied with the estimation of daily variables' effects. Indeed with our spline model we find a similar pattern across the hours as with model B and C. Model A simply smoothes out estimates, as intended. In this case, explanatory variables are not transformed and the different models estimate daily patterns as expected. Some improvement may however be required for the morning hours 6-7 on Saturdays and Sundays. Results for daily effects are overall consistent with operational models.

Regarding the impact of the modelling for the effect of hourly weather variables, we investigated more thoroughly forecasting accuracy for heating and non-heating periods. Although a small part of the forecast error variance can be attributed to an underestimation of the heating effect it does not explain the whole forecasting accuracy difference between model A on one side and models B and C on the other side. We also tried to impose larger values for the standard-deviations of the dynamics related to smoothed-heating to find again similar heating effects as in model C, but it had no important impact on overall forecasting accuracy.

We notice that the peak hours for which forecasting accuracy is poor are also hours for which the in-sample residual ACFs are worst, with a clear high intrayearly pattern. This explains an important part of the large forecast errors. We also investigated the orthogonal complement, $P'_\perp y_t$, of the reduced dimension vector y_t^*. It exhibits dynamic structure, showing that the dynamic factor $P'y_t$ does not capture all relevant dynamics. Consequently, forecasting this orthogonal complement as a white noise process is not satisfactory. More investigation is required to improve the model on this aspect.

We note however that model D with $R = 11$ knots already gives far more satisfactory results for non peak hours. Finally, we also point out in the discussion on parameter estimation results that hours with bad forecasting accuracy in Table 4.3 are associated with large estimated values in Σ_ε in Table 4.1. Poor in-sample explanatory power of λ_t for hour s corresponds to poor out-of-sample forecasting accuracy.

4.5 Conclusion

This chapter introduces intradaily smoothing splines in a dynamic regression model for hourly load. In particular, we discuss the periodic dynamic modelling of French national hourly electricity load for the period 1997-2007. The original time series is transformed into a vector of daily demands. Dynamic components include structural components as well as regression coefficients. The dimension of the dynamics for each stochastic component is reduced using cubic spline functions of the hour of the day. The loglikelihood of the data is evaluated in two steps. Following Jungbacker & Koopman (2008), the loglikelihood of the lower dimensional vector y_t^* is calculated via the prediction error decomposition. Then the result is adjusted to obtain the loglikelihood of the original time series y_t. We compute smoothed estimates of intradaily patterns of regression effects, and compare these with results for two models discussed in previous chapters. We also compare forecasts.

The application of the model for French hourly electricity demand is only a first step.

However we find the results encouraging, especially for the cooling effect and for the day-of-the-week effects. Our model specification is parsimonious compared to the previous dynamic modelling strategies. More components could still be added. For future research we consider different ways of empirical improvement for the spline model:

- Consider more knots (larger R) and/or knot positions;
- Incorporate dynamic factors for stochastic components where the dynamics dimension can be reduced, e.g. the stochastic trend;
- Considering a different number of knots and related position for different dynamic components;
- Improve model specification for weather-related variables;
- Consider linear splines (next to cubic splines).

4.6 Appendix: Spline weights calculations

This appendix describes the derivation of matrix W, as presented in Figure 4.2. Let $y_i = \theta(x_i)$, where $\theta(.)$ is a piecewise cubic spline function, interpolating to $y = \{y_0, \ldots, y_{R-1}\}$ using a set of knots $\{x_0^\dagger, \ldots, x_{R-1}^\dagger\}$. Imposing continuity and derivative conditions at interior knot positions $x_i^\dagger, i = 1, 2, \ldots, R-2$ leads to a linear system to solve, so that $\theta(.)$ should satisfy the following conditions, as detailed in Poirier (1973):

- The second derivative is piecewise linear. Defining $\theta''\left(x_i^\dagger\right) = a_i$ and $\theta''\left(x_{i-1}^\dagger\right) = a_{i-1}$ and rearranging terms gives for $x \in [x_{i-1}^\dagger; x_i^\dagger]$:

$$\theta''(x_i) = \frac{x_i^\dagger - x}{d_i} a_{i-1} + \frac{x - x_{i-1}^\dagger}{d_i} a_i, \quad d_i = x_i^\dagger - x_{i-1}^\dagger \tag{4.22}$$

- The first derivative is continuous giving:

$$\theta'\left(x_i^{\dagger+}\right) = \theta'\left(x_i^{\dagger-}\right) \tag{4.23}$$

In our case, the $x_i^\dagger$ are specific knot hours of the day. Suppose first that the cubic spline function is known for all knots: $\theta\left(x_i^\dagger\right) = y_i^\dagger$. Adding "natural" spline restrictions $a_0 = a_{R-1} = 0$ we obtain a linear system of R equations for the vector $a = (a_0 \ldots a_{R-1})'$. In matrix form: $\Lambda a = G y^\dagger$ so that $a = \Lambda^{-1} G y^\dagger$, where $y^\dagger = \left(y_0^\dagger \ldots y_{R-1}^\dagger\right)'$, and where Λ and G are $R \times R$ matrices with a tridiagonal structure. The equations follow from the resulting piecewise cubic expression for $\theta(x)$ presented in (4.24).

For any given position x,$x \in [x_{i-1}^{\dagger}; x_i^{\dagger}]$, the cubic spline function can be written as

$$\begin{aligned}\theta(x) = \; & \tfrac{x_i^{\dagger}-x}{6d_i}\left(\left(x_i^{\dagger}-x\right)^2 - d_i^2\right)a_{i-1} + \tfrac{x-x_{i-1}^{\dagger}}{6d_i}\left(\left(x-x_{i-1}^{\dagger}\right)^2 - d_i^2\right)a_i \\ & + \tfrac{x_i^{\dagger}-x}{d_i}y_{i-1}^{\dagger} + \tfrac{x-x_{i-1}^{\dagger}}{d_i}y_i^{\dagger}, \quad i = 1, \ldots, R-1.\end{aligned} \tag{4.24}$$

Again following Poirier (1973), we can easily evaluate $\theta(.)$ for an arbitrary S-vector x^*. In this chapter, $x^* = (0, 1, \ldots, S-1)$. The spline is expressed as $\theta\left(x_s\right) = w_s' y^{\dagger}$, where $w_s = (w_{s,1} \ldots w_{s,R})\prime$ where w_s does not depend on $y^{\dagger}$,$s = 1, \ldots, S$. Defining additional $(S \times R)$ matrices Q_1 and Q_2 from equation (4.24), Poirier (1973) derives the $(S \times R)$ matrix of spline weights W: $W = Q_1 \Lambda^{-1} G + Q_2$, which does not depend on $y^{\dagger}$.

Chapter 5

Conclusion and suggestions future research

This thesis investigates state-space modelling solutions in the case of high frequency data with specific applications to French national hourly electricity demand modelling and forecasting. Chapter 1 introduces the issues involved with short-term electricity load forecasting, describes intricacies of French load data and provides a short literature review on the topic. It also provides a concise introduction to state-space modelling with particular attention to applications in load forecasting. Chapters 2, 3, and 4 present three different new multivariate state-space models with detailed applications.

The dissertation thoroughly investigates multivariate state-space modelling but univariate models are also constructed as benchmark models. We first specify a periodic dynamic regression model adopting a vector structure for stochastic structural components such as trend and seasonals where we also consider time-varying regression coefficients. The model for electricity load typically involves many regressors. Since we let components vary over time, we estimate covariance matrices and we do not restrict these covariance matrices to be diagonal. The dynamic specifications for the time-varying components and for the coefficients involve covariance matrices relating the different equations of the dynamic regression model. We propose and estimate different specifications for these covariance matrices, and we study the practical implications for load modelling. Chapter 2 evaluates the benefit of multivariate models of electricity loads compared with univariate models.

The unrestricted specification involves large full rank matrices and practical estimation problems can occur, even for trivariate models for hourly electricity loads. We therefore propose a new specification with dynamic factors for the structural components and for the time-varying regression coefficients in chapter 3. For a successful estimation in the empirical study, we restrict the factor model to a block diagonal structure for the

factor loading matrices. This model specification is a first step towards the modelling of the intradaily dynamic pattern for each component of electricity demand. Finally we propose a third model specification where instead of estimating restricted factor loading matrices (and constant level adjustments) we use piecewise cubic splines to model the daily pattern of the data in a smooth way in chapter 4. For simplicity, we impose the spline specification, number of knots and knot positions, so that factor loadings are fixed a priori. The spline knots are modelled using periodic dynamic regression models. In this way we manage to estimate a model for the complete daily curve which is more parsimonious in the number of parameters to estimate.

In short, we propose different complementary specifications for high-frequency non-stationary data including stochastic regression effects on exogenous variables. Models are evaluated both from the signal extraction and post-sample forecasting accuracy standpoints.

We find several interesting empirical results for modelling and forecasting French hourly electricity loads. Benchmark univariate models already give us new satisfactory insights: the dynamic regression framework captures slowly evolving components such as the cooling effect in the summer but also highly time-varying features such as the strong variation of the yearly pattern during the summer vacations. The methodology is flexible enough to capture these different phenomena.

The recurrent non-linear shape of the time-variation of the heating regression coefficient in the winter is also an interesting empirical finding: the question of an intermediate heating coefficient at the beginning and at the end of the heating period is indeed discussed at EDF and our empirical results suggest a modelling possibility. The changes in the heating effect should become even more and more important in the future for the forecasting of the EDF-signal now that individual customers (these customers use a large part of electric heating) can leave EDF for a competitor. Future research shall focus on this intricate aspect of electricity demand.

The different multivariate models give consistent interpretations and show that most components have similar dynamics for the different hours. The spline model for the daily load curve in chapter 4 imposes smooth variations during the day. In most cases, we obtain estimated components consistent with the ones from EDF internal model.

The fact that the hyperparameters of the loglikelihood maximization are mostly standard deviations instead of regression coefficients themselves allows direct application of the specified models to new periods of the same data without deteriorating forecasting accuracy showing once again the flexibility of the state-space methodology, compared

to the internal model that needs specific adjustments for parameter extrapolation. This is illustrated with the benchmark models in chapter 4 that are directly applied to the extended dataset using hyperparameter estimates from chapter 3, in which the start of the estimation period is the same as in chapter 4 but the estimation sample is shorter, with no impact on forecasting accuracy.

The load decomposition between the different unobserved components where the components can have their own dynamics provides a useful analysis tool to understand total demand. The methodology also allows the immediate construction of realistic statistical forecast intervals. The empirical models estimate adaptive components that capture long-term evolution as well as strong short-term variations. We therefore constructed flexible models within the state-space framework that could be useful to forecast a portfolio demand in a more competitive environment, when applied to adequate data.

We suggest different further empirical research possibilities for the modelling and forecasting of national demand. Adequate model specification for special days (bank holidays and PDW days) should be investigated to build an operational model, especially for PDW days since they can occur theoretically 88 times a year, treating these days as missing leaves fewer days to estimate the heating effect: the forecasting for these days itself is also important for EDF. A further direction to improve our forecasting models is to modify the random-walk specification for the stochastic regression coefficients, especially for the heating effect (the specification in Chapter 3 was not completely satisfactory). This new specification allows to investigate the signal extraction but this should also be used to improve the forecasting specification for a weekly horizon. This extension can imply non-linear state-space modelling involving extensions of the standard Kalman filter recursions. The spline specification of chapter 4 can be extended to consider different spline specifications for different regression effects (different number of knots and/or knots positions). Periodic splines for a better modelling of the weekly seasonal can also be considered. The size of the vector to model can also be extended by modelling a weekly vector of hourly loads (i.e. of dimension 168).

In the situation of permanent changes in the electricity market involving customers repartition between the different market participants, portfolio variations are difficult to anticipate so that the time-varying dynamic regression framework can be useful to quickly adapt the different components of electricity demand. The multivariate approach can also improve robustness in this respect. We plan to specify a dynamic model for EDF portfolio data, which is a time-varying subset of national demand. However, for an efficient forecasting model of a time-varying customer portfolio, it would be preferable to get a priori information on some of the variations. If there is sufficient knowledge of

the portfolio constitution, it is possible to disaggregate the load curve to build separate models corresponding to homogeneous consumption behaviour. In this situation the random-walk specification may be sufficient to adjust the forecasts.

Another standpoint could be adopted with the specification of a dynamic model for internal mid-term model forecast errors to improve short-term forecasting. The approach would then be complementary with forecasting models in use at EDF.

Finally, electricity spot prices are strongly related to electricity demand since electricity storage is not possible. The models considered in this dissertation may be a basis for this topic.

Bibliography

Al-Hamadi, H. M. & Soliman, S. A. (2004), 'Short-term electric load forecasting based on kalman filtering algorithm with moving window weather and load model', *Electric Power Systems Research* **68**, 47–59.

Alfares, H. K. & Nazeeruddin, M. (2002), 'Electric load forecasting: literature survey and classification of methods', *International Journal of Systems Science* **33**, 23–34.

Alonso, A. M., Garcia-Martos, C., Rodriguez, J. & Sanchez, M. J. (2008), Seasonal dynamic factor analysis and bootstrap inference: Application to electricity market forecasting, Technical report, Universidad Carlos III, Departamento de Estadística y Econometría, Getafe, Spain.

Ansley, C. F. & Kohn, R. (1985), 'Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions', *Annals of statistics* **13**, 1286–1316.

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**, 191–221.

Bowsher, C. G. & Meeks, R. (2008), 'The dynamics of economic functions: Modeling and forecasting the yield curve', *Journal of the American Statistical Association* **103**, 1419–1437.

Box, G. E. P. & Jenkins, G. M. (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, San-Francisco, CA, USA.

Bruhns, A., Deurveilher, G. & Roy, J. S. (2005), A non linear regression model for mid-term load forecasting and improvements in seasonality, *in* 'Proceedings of the 15th Power Systems Computation Conference 2005, Liege Belgium'.

Bunn, D. W. & Farmer, E. D., eds (1985), *Comparative Models for Electrical Load Forecasting*, John Wiley, New York.

Cancelo, J. R. & Espasa, A. (1996), 'Modelling and forecasting daily series of electricity demand', *Investigaciones economicas* **20**, 359–376.

Cancelo, J. R., Espasa, A. & Grafe, R. (2008), 'Forecasting the electricity load from one day to one week ahead for the spanish system operator', *International journal of forecasting* **24**, 588–602.

Contaxi, E., Delkis, C., Kavatza, S. & Vournas, C. (2006), The effect of humidity in a weather-sensitive peak load forecasting model, *in* 'Proceedings of the 2006 Power Systems Conference and Exposition (PSCE) conference', IEEE Power Engineering Society, pp. 1528–1533.

Cottet, R. & Smith, M. (2003), 'Bayesian modeling and forecasting of intraday electricity load', *Journal of the American Statistical Association* **98**, 839–849.

De Gooijer, J. G. & Hyndman, R. J. (2006), '25 years of time series forecasting', *International Journal of Forecasting* **22**, 443–473.

De Gooijer, J. G. & Ray, B. K. (2003), 'Modeling vector nonlinear time series using POLYMARS', *Computational Statistics and Data Analysis* **42**, 73–90.

De Jong, P. (1991), 'The diffuse Kalman filter', *Annals of Statistics* **19**, 1073–1083.

Del Negro, M. & Otrok, C. (2008), Dynamic factor models with time-varying parameters: measuring changes in international business cycles, Technical report, Federal Reserve Bank of New York, New York, USA.

Doornik, J. A. (2006), *Ox 5 : An Object-oriented Matrix Programming language*, Timberlake Consultants Ltd, www.oxmetrics.com, London, U.K.

Dordonnat, V., Koopman, S. J. & Ooms, M. (2009), Dynamic factors in state-space models for hourly electricity load signal decomposition and forecasting, *in* 'Proceedings of the IEEE PES General meeting 2009,Calgary, Canada.'.

Dordonnat, V., Koopman, S. J., Ooms, M., Collet, J. & Dessertaine, A. (2008), 'An hourly periodic state space model for modelling french national electricity load', *International Journal of Forecasting* **24**, 566–587.

Durbin, J. & Koopman, S. J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, UK.

Engle, R. F., Granger, C. W. J., Rice, J. & Weiss, A. (1986), 'Semiparametric estimates of the relation between weather and electricity', *Journal of the American Statistical Association* **81**, 310–320.

Engle, R. & Watson, M. W. (1981), 'A one-factor multivariate time series model of metropolitan wage rates', *Journal of the American Statistical Association* **76**, 774–781.

Espinoza, M., Joye, C., Belmans, R. & De Moor, B. (2005), 'Short-term load forecasting, profile identification and customer segmentation : A methodology based on periodic time series', *IEEE Transactions on Power Systems* **20**(3), 1622–1630.

Fan, J. Y. & Mc Donald, J. D. (1993), 'A real-time implementation of short-term load forecasting for distribution power systems', *IEEE Transactions on Power Systems* **9**, 988–994.

Fletcher, R. (1987), *Practical Methods of Optimisation, (2nd Ed.)*, John Wiley, New York.

Francke, M. (2006), Marginal likelihood in State space models; Theory and applications, PhD thesis, Vrije Universiteit Amsterdam, Netherlands.

Francke, M., Koopman, S. J. & de Vos, A. (2008), Likelihood functions for state-space models with diffuse initial conditions, Technical report, Tinbergen institute discussion paper, The Netherlands.

Gastaldi, M., Lamedica, R., Nardecchia, A. & Prudenzi, A. (2004), Short-term forecasting of municipal load through a kalman filtering based approach, *in* 'Proceedings of the 2004 Power Systems Conference and Exposition (PSCE) conference', IEEE Power Engineering Society, pp. 1453–1458.

Giannone, D., Reichlin, L. & Sala, L. (2006), 'VARs, common factors and the empirical validation of equilibrium business cycle models', *Journal of Econometrics* **132**, 257–279.

Goude, Y. (2008), Melange de predicteurs, application a la prevision de consommation d'electricite, PhD thesis.

Harvey, A. C. (1989), *Forecasting, structural time series models and the Kalman Filter*, Cambridge University Press, Cambridge, UK.

Harvey, A. C. & Koopman, S. J. (1993), 'Forecasting hourly electricity demand using time-varying splines', *Journal of the American Statistical Association* **88**, 1228–1237.

Harvey, A. C. & Koopman, S. J. (1997), Multivariate structural time series models, *in* C. Heij, H. Schumacher, B. Hanzon & C. Praagman, eds, 'Systematic Dynamics in Economic and Financial Models', Wiley, Chichester, UK., pp. 269–298.

Hippert, H. S., Bunn, D. W. & Souza, R. W. (2005), 'Large neural networks for electricity load forecasting: Are they overfitted?', *International Journal of Forecasting* **21**, 425–434.

Hyndman, R. J. & Fan, S. (2008), Density forecasting for long-term peak electricity demand, Technical report, Department of econometrics and business statistics, Monash University, A.

Jungbacker, B. & Koopman, S. J. (2008), Likelihood-based analysis for dynamic factor models, Technical report, Tinbergen institute discussion paper, The Netherlands.

Koopman, S. J. (1993), 'Disturbance smoother for state space models', *Biometrika* **80**, 117–126.

Koopman, S. J. (1997), 'Exact initial kalman filtering and smoothing for non-stationary time series models', *Journal of the American Statistical Association* **92**, 1630–1638.

Koopman, S. J. & Durbin, J. (2000), 'Fast filtering and smoothing for multivariate state space models', *Journal of Time Series Analysis* **21**, 281–296.

Koopman, S. J. & Ooms, M. (2003), 'Time series modelling of daily tax revenues', *Statistica Neerlandica* **57**, 439–469.

Koopman, S. J. & Ooms, M. (2006), 'Forecasting daily time series using periodic unobserved components time series models', *Computational Statistics & Data Analysis* **51**, 885–903.

Koopman, S. J. & Shephard, N. (1992), 'Exact score for time series models in state space form', *Biometrika* **79**, 823–826.

Koopman, S. J., Shephard, N. & Doornik, J. A. (1999), 'Statistical algorithms for models in state space using SsfPack 2.2', *The Econometrics Journal* **2**, 107–160, www.ssfpack.com.

Koopman, S. J., Shephard, N. & Doornik, J. A. (2008), *SsfPack 3.0*, Timberlake Consultants Ltd, London, U.K.

Laluque, N. (2007), A generalized additive model for mid-term french load forecasting and introduction of wind speed, *in* 'Power Systems Modelling Conference', Athens, Greece.

Lefieux, V. (2007), Modeles semi-parametriques appliques a la prevision des series temporelles - Cas de la consommation d'electricite, PhD thesis.

Liu, J. M., Chen, R., Liu, L.-M. & Harris, J. L. (2006), 'A semi-parametric time series approach in modeling hourly electricity loads', *Journal of Forecasting* **25**, 537–559.

Lotufo, A. D. P. & Minussi, C. R. (1999), Electric power systems load forecasting: A survey, *in* 'IEEE Power Tech Conference, Budapest Hungary'.

Martin, M. M. (1999), 'Filtrage de Kalman d'une serie temporelle saisonniere - application a la prevision de consommation d'electricite', *Revue de Statistiques appliquees* pp. 69–86.

Martin-Rodriguez, G. & Caceres-Hernandez, J. J. (2005), 'Modelling the hourly Spanish electricity demand', *Economic Modelling* **22**, 551–569.

Ménage, J. P., Panciatici, P. & Boury, F. (1988), Nouvelle modélisation de l'influence des conditions climatiques sur la consommation d' énergie électrique, Technical Report HR-25/2164, EDF R&D.

Misiti, M., Misiti, Y., Oppenheim, G. & Poggi, J.-M. (2008), Optimized clusters for disaggregated electricity load forecasting, *in* 'Proceedings of COMPSTAT', ISI, Porto, Portugal, p. paper 139.

Muñoz Carpena, R., Ritter, A. & Li, Y. C. (2005), 'Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to everglades national park', *Journal of Contaminant Hydrology* **80**, 49–70.

Ortega, J. A. & Poncela, P. (2005), 'Joint forecasts of southern european fertility rates with non-stationary dynamic factor models', *International Journal of Forecasting* **21**, 539–550.

Pai, P.-F. & Hong, W.-C. (2005), 'Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms', *Electric Power Systems Research* **74**, 417–425.

Pardo, A., Meneua, V. & Valor, E. (2002), 'Temperature and seasonality influences on spanish electricity load', *Energy economics* **24**, 55–70.

Pedregal, D. J. & Young, P. C. (2006), 'Modulated cycles, an approach to modelling periodic components from rapidly sampled data', *International Journal of Forecasting* **22**, 181–194.

Pedregal, D. J. & Young, P. C. (2008), 'Development of improved adaptive approaches to electricity demand forecasting', *Journal of the Operational Research Society* **59**, 1066–1076.

Peirson, J. & Henley, A. (1994), 'Electricity load and temperature', *Energy economics* **16**, 235–243.

Peña, D. & Poncela, P. (2004), 'Forecasting with nonstationary dynamic factor models', *Journal of Econometrics* **119**, 291–321.

Poggi, J.-M. (1994), 'Prévision non paramétrique de la consommation électrique', *Revue de statistique appliquée* pp. 83–98.

Poirier, D. J. (1973), 'Piecewise regression using cubic spline', *Journal of the American Statistical Association* **68**, 515–524.

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F. & Brace, C. (1997), 'Short-run forecasts of electricity loads and peaks', *International Journal of Forecasting* **13**, 161–174.

Shephard, N. (1993), 'Distribution of the ml estimator of an MA(1) and a local level model', *Econometric theory* **9**, 377–401.

Shumway, R. H. & Stoffer, D. S. (1982), 'An approach to time series smoothing and forecasting using the EM algorithm', *Journal of Time Series Analysis* **3**, 253–263.

Sisworahardjo, N. S., El-Keib, A. A., Choi, J., Valenzuela, J., Brooks, R. & El-Agtal, I. (2006), 'A stochastic load model for an electricity market', *Electric Power Systems Research* **76**, 500–508.

Smith, M. (2000), 'Modeling and short-term forecasting of new south wales electricity system load', *Journal of Business & Economic Statistics* **18**, 465–478.

Smith, M. & Kohn, R. (2002), 'Parsimonious covariance matrix estimation for longitudinal data', *Journal of the American Statistical Association* **97**, 1141–1153.

Soares, L. J. & Medeiros, M. C. (2008), 'Modeling and forecasting short-term electricity load: a comparison of methods with an application to brazilian data', *International Journal of Forecasting* **24**, 630–644.

Soares, L. J. & Souza, L. R. (2006), 'Forecasting electricity demand using generalized long memory', *International Journal of Forecasting* **22**, 17–28.

Stock, J. H. & Watson, M. W. (1998), 'Median unbiased estimation of coefficient variance in a time-varying parameter model', *Journal of the American Statistical Association* **93**, 349–358.

Taylor, J. W. (2003), 'Short-term electricity demand forecasting using double seasonal exponential smoothing', *Journal of the Operational Research Society* **54**, 799–805.

Taylor, J. W. (2008), 'An evaluation of methods for very short-term load forecasting using minute-by-minute british data', *International Journal of Forecasting* **24**, 645–658.

Taylor, J. W. & Buizza, R. (2003), 'Using weather ensemble predictions in electricity demand forecasting', *International Journal of Forecasting* **19**, 57–70.

Taylor, J. W., De Menezes, L. M. & McSharry, P. E. (2006), 'A comparison of univariate methods for forecasting electricity demand up to a day ahead', *International Journal of Forecasting* **22**, 1–16.

Taylor, J. W. & McSharry, P. E. (2007), 'Short-term load forecasting methods: An evaluation based on european data', *IEEE Transactions on Power Systems* **22**, 2213 – 2219.

Tiao, G. C. & Grupe, M. R. (1980), 'Hidden periodic autoregressive-moving average models in time series data', *Biometrika* **67**, 365–373.

Tunnicliffe Wilson, G. (1989), 'On the use of marginal likelihood in time series model estimation', *Journal of the Royal Statistical Society, Series B* **51**, 15–27.

Wang, B., Neng-ling, T., Hai-qing, Z., Jian, Y., Jia-dong, Z. & Liang-bo, Q. (2008), 'A new ARMAX model based on evolutionary algorithm and particle swarm optimization for short-term load forecasting', *Electric Power Systems Research* **78**, 1679–1685.

Watson, M. W. & Engle, R. F. (1983), 'Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression models', *Journal of Econometrics* **23**, 385–400.

Young, P. C., Pedregal, D. J. & Tych, W. (1999), 'Dynamic harmonic regression', *Journal of Forecasting* **18**, 369–394.

Zheng, T., Girgis, A. A. & Makram, E. B. (2000), 'A hybrid wavelet-kalman filter method for load forecasting', *Electric Power Systems research* **54**, 11–17.

Summary

This dissertation explores successively three models specifications within the multivariate linear Gaussian state-space framework for an application to high-frequency data such as hourly electricity demand. Empirical results are obtained for French national data. The motivation is the need for Electricité De France (EDF) to ivestigate new statistical methodologies for electricity demand forecasting in order to face new market conditions.

First the short-term load forecasting issue and a state of the art in demand forecasting are presented. The state-space methodology is introduced. The three models presented in the dissertation are designed for a vector time series of data.

The first modelling approach is a periodic time-varying regression model where the dynamics of each specific component can be correlated for different hours of the day. The univariate and succesive bivariate models for French hourly demand give a good one-day-ahead forecasting accuracy and interesting features in the decomposition of demand between seasonals and weather effects are discussed.

The second approach present a restriction of the previous model using dynamic factors in the trend as well as in time-varying regression coefficients. The model has a block diagonal structure for common structures so that trivariate models are estimated.

The third approach focuses on the intradaily pattern for all components of electricity demand. Piecewise cubic splines are used in this respect in order to smooth the factor structure of the second model. The result is a very parsimonious model that gives satisfactory insights for the seasonal pattern.

With these models, we manage to estimate highly intrayearly time-varying features (summer holidays) as well as slowly evolving features (air conditioning effect). We conclude that the methodology is flexible enough to adapt to variations in the customer portfolio, offering new model possibilities for EDF. Some suggestions for future research are given in the final chapter.

Nederlandse samenvatting

Dit proefschrift met de title "Toestandsruimtemodellering van hoogfrequente data, drie toepassingen voor de Franse nationale vraag naar elektriciteit" ontwikkelt drie dynamische modellen voor hoogfrequente tijdreeksdata, in het bijzonder voor de vraag naar elektriciteit per uur van de dag. Deze drie toestandsruimtemodellen zijn meerdimensionaal en combineren latente variabelen voor trends en seizoenseffecten met waarneembare verklarende variabelen zoals de temperatuur. Het onderzoek is gemotiveerd door de behoefte aan nieuwe statistische methoden bij Electricité de France (EDF) voor het voorspellen van de elektriciteitsvraag onder veranderende marktomstandigheden. Het empirische onderzoek van deze dissertatie heeft diverse nieuwe inzichten opgeleverd over het verloop van de nationale France vraag naar elektriciteit in de periode 1997-2007.

Het proefschrift bevat vijf hoofdstukken. Het eerste hoofdstuk behandelt de problemen bij het voorspellen van de elektriciteitsvraag en bespreekt kort de laatste ontwikkelingen op dit gebied uit de literatuur. Toestandsruimtemodellen (State-space models) en-methoden worden geïntroduceerd.

Hoofdstuk 2 presenteert ons eerste nieuwe model. Hierin beschouwen we een meerdimensionaal periodiek regressiemodel met stochastische tijdsafhankelijke parameters. Het model verklaart een vraagvector voor de verschillen uren van de dag. Het model houdt expliciet rekening met correlaties tussen veranderingen tussen vergelijkbare latente variabelen en coëfficiënten voor de verschillende uren. We schatten een tweedimensionale versie van het model voor verschillende combinaties van uren en passen het model ook toe voor alle uren afzonderlijk. De één- en tweedimensionale modellen vertonen een goede voorspelprecisie voor een voorspelhorizon van één dag bij toepassing op de lange tijdreeks van de Franse elektriciteitsvraag. Bovendien levert het model interessante en goed te onderscheiden patronen op voor de algemene seizoeneffecten en voor de specifieke weereffecten in de vraag. Het derde hoofdstuk behandelt een nieuwe variant van het model uit hoofdstuk 2. Hierin leggen we nieuwe beperkingen op aan de correlatiestructuur die de stochastische veranderingen in de modelvergelijkingen voor de verschillende uren van de dag met elkaar verbindt. We gebruiken verschillende dynamische-factor-specificaties

voor de latente variabelen en voor de tijdvariërende regressiecoëfficiënten. In de toepassing op de 24 uren met dagelijkse data van de Franse vraag gebruiken we een diagonale correlatiestructuur met blokken voor telkens drie opeenvolgende uren.

Hoofdstuk 4 introduceert onze derde specificatie van een meerdimensionaal periodiek toestandsruimtemodel voor hoogfrequente data. De totale daglijkse elektriciteitsbelastingcurve van Frankrijk wordt hierin direct gemodelleerd als een gladde combinatie van lokale derdegraadsfuncties van het uur van de dag. Dit model is relatief spaarzaam geparameteriseerd en geeft een goed inzicht in de trendmatige veranderingen in de dagelijkse en wekelijkse belastingcurve in de loop van het jaar en in de loop der jaren. In hoofdstuk 4 maken we ook een empirische vergelijking tussen de drie modellen van het proefschrift.

De nieuw ontwikkelde modellen stellen ons in staat om veel verschillende effecten tegelijkertijd te meten, zoals grote en abrupte veranderingen in de vraag door de zomervakantie en subtiele trendmatige veranderingen als het effect van airconditioning die we alleen kunnen waarnemen bij hoge temperaturen.

In het afsluitende hoofdstuk stellen we vast dat onze methodologie voldoende veelomvattend en flexibel is voor vraagmodellering van grote portefeuilles met sterk veranderende klantenbestanden. Onze benadering biedt goede mogelijkheden voor toekomstig gebruik bij EDF.